Scenario generation quality assessment for two-stage stochastic programs

by

Didem Sari

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial and Manufacturing Systems Engineering

Program of Study Committee:
Sarah M. Ryan, Major Professor
Guiping Hu
Kyung J. Min
Li Wang
Lizhi Wang

Iowa State University

Ames, Iowa

2017

DEDICATION

I would like to dedicate this dissertation to my beloved husband and best friend Dr. Muhammet Ay

TABLE OF CONTENTS

LIST OF TABLES vi

LIST OF FIGURES vii

ACKNOWLEDGEMENTS xii

ABSTRACT xiii

CHAPTER 1. INTRODUCTION 1

1.1 Background 1

1.2 Motivation 2

1.3 Problem Statement 4

1.4 Thesis Structure and Overview 4

References 6

CHAPTER 2. STATICTICAL METRICS FOR ASSESSING THE QUALITY OF WIND POWER SCENARIOS FOR STOCHASTIC UNIT COMMITMENT 8

Abstract 8

2.1 Introduction 9

2.2 Verification of scenarios 13

2.2.1 Energy scores 14

2.2.2 Distance-based rank histograms 15

2.2.3 Event-based verification 32

2.3 Wind power scenario generation methods 34

2.3.1 Wind power scenario generation by quantile regression with Gaussian copula approach 36

2.3.2 Wind power scenario generation by epi-spline approximation approach 40

2.4 Example application of the verification approaches 43

2.4.1 The BPA dataset 43

2.4.2 Verification of BPA scenarios 45

2.5 Conclusions 54

References 55

CHAPTER 3. RELIABILITY OF WIND POWER SCENARIOS AND STOCHASTIC UNIT COMMITMENT COST 60

Abstract 60

3.1 Introduction 61

3.2 Literature Review 64

3.3 Statistical metrics for wind power scenario evaluation 67

3.3.1 Mass transportation distance rank histogram 69

3.3.2 Event-based verification 72

3.4 Stochastic unit commitment and dispatch problem 74

3.4.1 Simulation procedure over a rolling horizon 76

3.5 Case study 78

3.5.1 The dataset 78

3.5.2 Wind power scenario generation 79

3.5.3 Results 80

3.5.3.1 SUC and dispatch simulation results 81

3.5.3.2 Statistical metric results 85

3.5.4 Discussion 88

3.5.5 Daily comparisons 91

3.6 Conclusions 95

References 96

CHAPTER 4. SCENARIO GENERATION ASSESSMENT FOR TWO-STAGE STOCHASTIC PROGRAMS 102

Abstract 102

4.1 Introduction 103

4.2 Literature Review 106

4.3 Scenario quality assessment by historical simulation 110

4.4 Proposed generic approach for assessment of scenario generation methods 112

4.5 Motivating models 116

4.6 Expected value based scenario assessment 118

4.7 Perfect information based scenario assessment 122

4.7.1 Simulation study on server location problem 123

4.7.2 Simulation study on unit commitment problem 124

4.7.3 Evaluation of MTD rank histogram for PI-based scenario assessment 130

4.8 Case study of EV-based and PI-based scenario assessment – unit commitment problem 132

4.9 Conclusions 138

References 139

CHAPTER 5. GENERAL CONCLUSION 143

LIST OF TABLES

Table 2.1 Overview of scenario generation approaches 35

Table 2.2 Means and standard deviations of energy scores for scenarios generated by quantile regression with Gaussian copula approach with various combinations of models and nonlinearity 49

Table 2.3 Means and standard deviations of energy scores for epi-spline approximation approach scenarios with different cutting probabilities 49

Table 2.4 Average daily Brier scores for ramp down and ramp up events with magnitudes ξ = 0.2, 0.4 and durations κ = 1, 3, 6 for scenarios generated by two variants of the epi-spline approximation approach and by three variants of the quantile regression with Gaussian copula approach 53

Table 3.1 Date IDs and date ranges of the 4 historical time periods used for the SUC simulations 81

Table 3.2 Percentages of deviations from the optimal generation levels by Epi spline scenarios 83

Table 3.3 Percentages of deviations from the optimal generation levels by quantile regression scenarios 85

Table 3.4 Average daily Brier scores for RampDown event with two different parameters for scenarios EPI(0.1, 0.9), EPI(0.33, 0.66), QR, and QRnew 87

LIST OF FIGURES

Figure 2.1 A hypothetical example in 2 dimensions is presented for minimum spanning tree and mass transportation distance 19

Figure 2.2 The observation 0, is moved to point (9,7) from point (4,2) to obtain an under-dispersed and/or biased ensemble, whereas the 8 scenarios labeled 1-8, have been kept at their same coordinates 20

Figure 2.3 Minimum spanning tree rank histogram – simulation study 22

Figure 2.4 Mass transportation distance rank histogram – simulation study with the same setup as in Fig. 2.3 23

Figure 2.5 MTD and MST rank histograms for over-dispersed scenarios with and without bias 24

Figure 2.6 Minimum spanning tree rank histogram – simulation study for testing scenarios according to their autocorrelations 27

Figure 2.7 Mass transportation distance rank histogram – simulation study for testing scenarios according to their autocorrelations with the same setup as in Fig. 2.6 28

Figure 2.8 MTD rank histograms for $\sigma_{obs}/\sigma_{scen} = 1$ and various combinations of ρ_{obs} and ρ_{scen} 29

Figure 2.9 MTD rank histograms when $\sigma_{obs}/\sigma_{scen} = 1$ and scenarios with both $\rho_{scen} > \rho_{obs}$ and $\rho_{scen} < \rho_{obs}$ are present 30

Figure 2.10 MTD and MST rank histograms 31

Figure 2.11 Wind power scenarios generated for day 2012-11-08 46

Figure 2.12 Wind power scenarios generated for day 2013-03-26. Note the difference in scale between the left- and right-hand panels 47

Figure 2.13 Wind power scenarios generated for day 2013-04-08 48

Figure 2.14 Mass transportation distance rank histograms for scenarios generated by (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) QR(1,1) and (d) QR(4,3) 51

Figure 2.15 (a) MST rank histogram when the scenario probabilities are not considered, (b) MTD rank histogram incorporating scenario probabilities 52

Figure 3.1 SUC and dispatch simulation procedure over a historical time period 76

Figure 3.2 Cumulative SUC and dispatch costs over 4 historical time periods 82

Figure 3.3 Cumulative deviations from optimal generation level occurred by epi-spline scenarios over 4 historical time periods 83

Figure 3.4 Cumulative deviations from optimal generation levels occurred by quantile regression scenarios over 4 planning historical time periods 84

Figure 3.5 Mass transportation distance rank histograms for scenarios EPI(0.1, 0.9), EPI(0.33, 0.66), QR, and QRnew 86

Figure 3.6 Hourly average load and average hourly Brier scores for Event 1 and 2 87

Figure 3.7 Wind power scenarios generated for day 2012/10/19. (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) QR, (d) QRnew 92

Figure 3.8 Wind power scenarios generated for day 2013/04/19. (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) QR, (d) QRnew 93

Figure 3.9 Wind power scenarios and net load scenarios for day 2013/04/14. (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) Net load scenarios for EPI(0.1, 0.9), (d) Net load scenarios for EPI(.33, 0.66) 94

Figure 4.1 Generic scenario assessment 116

Figure 4.2 EV-based scenario assessment 119

Figure 4.3 MTD rank histograms obtained by EV-based scenario assessment for various values of parameter $p_{scen} = (\mathrm{a})\ 0.1\ (\mathrm{b})\ 0.3\ (\mathrm{c})\ 0.4\ (\mathrm{d})\ 0.5\ (\mathrm{e})\ 0.6\ (\mathrm{f})\ 0.7\ (\mathrm{g})\ 0.9$ when $\mathrm{p}_{obs} = 0.5$. 120

Figure 4.4 MTD rank histograms obtained by expected value based scenario assessment for various values of parameter $\rho_{scen} = (\mathrm{a})\ 0.01^2\ (\mathrm{b})\ 0.05^2\ (\mathrm{c})\ 0.1^2\ (\mathrm{d})\ 0.2^2\ (\mathrm{e})\ 0.6^2\ (\mathrm{f})\ 0.9^2$ when $\rho_{obs} = 0.1^2$. 121

Figure 4.5 PI-based scenario assessment 122

Figure 4.6 MTD rank histograms obtained by PI-based scenario assessment for various values of parameter $p_{scen} = (\mathrm{a})\ 0.1\ (\mathrm{b})\ 0.3\ (\mathrm{c})\ 0.4\ (\mathrm{d})\ 0.5\ (\mathrm{e})\ 0.6\ (\mathrm{f})\ 0.7$ when $p_{obs} = 0.5$ 123

Figure 4.7 MTD rank histograms obtained by PI-based scenario assessment for various values of parameter $\rho_{scen} = (a)\ 0.01^2\ (b)\ 0.05^2\ (c)\ 0.1^2\ (d)\ 0.2^2\ (e)\ 0.6^2\ (f)\ 0.9^2$ when $\rho_{obs} = 0.1^2$. 123

Figure 4.8 MTD rank histograms obtained by PI-based scenario assessment for $\mu_{obs} = \mu_{scen} = 2500,\ \sigma_{obs} = \sigma_{scen} = 100,$ and $\rho_{obs} = \rho_{scen}.70$. 125

Figure 4.9 Possible MTD rank histograms obtained by PI-based scenario assessment when we try different combinations of parameters while keeping $\mu_{obs} = \mu_{scen},\ \sigma_{obs} = \sigma_{scen},$ and $\rho_{obs} = \rho_{scen}$ (a)$(\sigma_{obs} = 70, \mu_{obs} = 2500, \rho_{obs} = .75)$, $(b)(\sigma_{obs} = 300, \mu_{obs} = 1500, \rho_{obs} = .50)$, $(c)(\sigma_{obs} = 800, \mu_{obs} = 800, \rho_{obs} = .50)$. 126

Figure 4.10 MTD rank histograms obtained by PI-based scenario assessment when applied to unit commitment problem for various values of parameters $\mu_{scen} = (a)\ 1500\ (b)\ 2000\ (c)\ 2750\ (d)\ 3000.$ 127

Figure 4.11 MTD rank histograms obtained by PI based scenario assessment when applied to unit commitment problem for various values of parameter $\frac{\sigma_{obs}}{\sigma_{scen}} = (a)\ 4\ (b)\ 2\ (c)\ \frac{1}{2}\ (d)\ \frac{1}{4}.$ 128

Figure 4.12 MTD rank histogram obtained by PI-based scenario assessment for $\mu_{scen} = 2000$, $\frac{\sigma_{obs}}{\sigma_{scen}} = \frac{1}{4}$, and $\rho_{obs} = \rho_{scen}$. 129

Figure 4.13 MTD rank histograms obtained by PI based scenario assessment for heterogeneous ρ_{scen} when applied to unit commitment problem, where ρ_{scen} consists of (a) $(.70,\ .85,\ .95)$, (b) $(.70,\ .75,\ .80)$, (c) $(.70,\ .60,\ .50)$, (d) $(.60,\ .40,\ .20)$. 130

Figure 4.14 Random scenario based scenario assessment 131

Figure 4.15 MTD rank histogram obtained by a randomly selected scenario to evaluate the MTD rank histogram shown in Fig. 4.8 132

Figure 4.16 Wind power scenarios generated for the same day. (a) EPIwide, (b) EPInarrow, (c) QR, (d) QRnew 133

Figure 4.17 MTD rank histograms obtained by EV-based scenario assessment when applied to UC problem with scenarios generated by approaches (a) EPIwide, (b) EPInarrow, (c) QR and (d) QRnew. 134

Figure 4.18 MTD rank histograms obtained with scenarios generated by approaches EPIwide and QRnew by (a) PI-based scenario assessment and (b) randomly selected scenario for assessment. 136

ACKNOWLEDGEMENTS

I would like to gratefully thank my major professor Dr. Sarah M. Ryan for her excellent guidance, patience, and encouragement throughout my graduate study. Her profound knowledge in the field, constructive comments and advices have been of great value for me and were critically important for completing this dissertation. I would also like to thank my committee members, Guiping Hu, Jo Min, Lily Wang, and Lizhi Wang for their help and comments to improve the quality of my research.

I also want to acknowledge the Advanced Research Projects Agency - Energy (ARPA-E) and Iowa Energy Center for the funding support.

Finally, my deepest gratitude goes to my husband Dr. Muhammet Ay and my parents for their unconditional love, encouragement, and support, without which I would not have been able to finish my PhD.

ABSTRACT

In minimization problems with uncertain parameters, cost savings can be achieved by solving stochastic programming (SP) formulations instead of using expected parameter values in a deterministic formulation. To obtain such savings, it is crucial to employ high quality probabilistic scenarios for the uncertain parameters. A convincing way to assess the quality of a scenario generation method is to simulate employing the resulting scenarios when solving the SP problem while measuring the costs incurred when the solution is implemented and observed parameter values occur. Simulation studies to assess the quality in this way are computationally very demanding. This research is aimed at developing faster methods to assess the quality via statistical metrics. Relibility, which is defined as the statistical consistency between scenarios and observation, is a prerequisite for quality. The dissertation is presented in a three-paper format.

The stochastic unit commitment problem in electric power system operation is an application of SP that motivated this study. In power systems with high penetration of wind generation, probabilistic scenarios for the available wind energy are generated for use in stochastic formulations of day-ahead thermal unit commitment problems. To minimize the expected cost of dispatching the committed units, the wind energy scenarios should accurately represent the stochastic process for available wind energy. In the first paper, aiming to assess the reliability of probabilistic scenarios for wind energy time series, we employ some existing forecast verification approaches and introduce a mass transportation distance rank histogram to assess the reliability of unequally likely scenarios. In the second paper, we examine the relationship between the statistical reliability assessment metrics and the cost results of solving

SUC using the assessed scenario generation method. Based on this relationship, we understand the importance of scenario reliability to ensure scenario quality for the SUC problem. In the third paper, we extend this work to make it more robust and general enough to be applied to any two-stage SP problem that is repeatedly solved and for which observational data exist for some historical period. Focusing on scenario quality, we develop two novel approaches: expected value based and perfect information based scenario generation assessment. With the proposed approaches, we can assess the quality of scenario sets without having to repeatedly solve the related SP problem. Instead of comparing scenarios to observations directly, these approaches take into consideration the impact of each scenario on the solution to the SP problem.

CHAPTER 1: INTRODUCTION

1.1 Background

A stochastic program is a mathematical program in which some of the parameters are modeled as random variables [1]. In recent years, stochastic programming has gained an increasing popularity. Stochasticity can be added to optimization models by describing their uncertain parameters in terms of probability distributions instead of specifying single values. In the two stage stochastic programming approach, the decision variables are partitioned into two sets: the first stage and the second stage, or recourse, variables. The first stage variables must be decided before the actual realization of the uncertain parameters becomes available. The second stage variables are decided after the random events underlying the uncertain parameters occur. The objective is to choose the first stage variables such that the sum of first stage costs and the expected value of the random second stage costs is minimized.

In the deterministic equivalent formulations that can be solved, discrete probabilistic scenarios are employed to approximate the joint distribution of the uncertain parameters. To achieve a good solution to the stochastic program we must formulate a finite number of reasonable scenarios. The quality of the stochastic programming solution is directly linked to the quality of the scenarios. There is a huge literature on stochastic programming with contributions to the modeling and computational aspects. Within scenario generation approaches, different variants and parameter settings can produce different sets of scenarios. Although computational methods for solving stochastic programming problems have improved significantly, simulating the performance of alternative scenario sets may still be computationally prohibitive.

1.2 Motivation

The stochastic unit commitment (SUC) problem [2] is an application of stochastic programming that motivated this study. As the wind energy industry grows, the penetration of wind in electric power systems deepens. Solving operational planning problems for electric power generation requires more sophistication to accommodate the increasing variability in supply due to wind and other sources of variable renewable generation. Unit commitment problems, which are optimization problems used to determine the operation schedule of each thermal generating unit over a given period of time, have traditionally accounted for uncertaintyin demand by imposing fixed reserve limits. With pre-determined reserve limits, the small amounts of variable generation could be managed; however, to accommodate the further integration of wind energy, the treatment of uncertainty must improve. To meet the challenge of operating the systems under the increased uncertainty due to the deep wind penetration, SUC formulations, in which the probabilistic scenarios represent the wind energy time series over the planning horizon, have been proposed and widely tested.

The SUC problem can be modeled as a two-stage stochastic program. Its aim is to identify a unit commitment schedule that minimizes the first stage costs as well as the expected second stage costs over all scenarios [2]. Commitments of units are decided in the first stage before information on load and variable resource availability is known, while decisions on dispatching the committed units in each period are delayed until the second stage after the information is known. The two-stage stochastic program minimizes startup and shutdown costs in the first stage as well as expected generation cost and penalties on load mismatch in the second stage while satisfying operational restrictions over all scenarios.

In SUC, implicit rather than fixed reserve limits are imposed by finding a unit commitment schedule that minimizes expected costs with respect to a set of probabilistic scenarios. Compared to the deterministic unit commitment, schedules meeting the load more cost-effectively can be obtained by solving the SUC, especially in systems with deep penetration of wind power [3, 4]. Costs are saved by committing more resources on days when the wind contribution has low expectation and/or high uncertainty and by committing fewer resources on days when the wind contribution has high expectation and/or low uncertainty.

Numerous studies of power system planning have been conducted to investigate how to find unit commitment solutions that effectively accommodate wind power uncertainty. Various technical approaches have been devised to determine optimal generation scheduling in a wind integrated power system. Previous SUC research has focused on improving the mathematical formulation, developing solution approaches to decrease the optimality gap, and devising various scenario reduction techniques to decrease the solution time. Various mathematical programming methods and optimization techniques have been applied to SUC problems, including branch-and-bound, dynamic programming, mixed-integer programming, and Lagrangian relaxation. Although an extensive literature exists on generating scenarios for SUC and other stochastic programs, the assessment of scenario generation methods according to the performance of the resulting scenario sets in the SUC problem is limited. Because the solution of a SUC problem is directly related to the employed scenarios, we aim to assess the quality of wind power scenarios.

1.3 Problem Statement

In this dissertation, our goal is to assess scenario sets – and, consequently, scenario generation methods that produce them – that are used in stochastic programming problems according to the associated objective function values. However, the challenging computational complexity of stochastic programs makes this judgment, according to the given criterion, cumbersome. The high computational demands of solving stochastic programs motivate a search for ways to evaluate the quality of scenario sets without extensively simulating the stochastic programming procedure, which requires the repeated solution of large deterministic equivalent mathematical programs.

We assume observational data are available for similar instances during some historical period and adopt a statistical approach. Because the scenarios directly affect the solution of the stochastic problem, it is crucial for the scenarios that would be generated for an instance to align with the observed data for that instance. Although advanced methods are applied to stochastic problems to reduce the computational effort, a simulation study to compare the results of using different scenario sets remains computationally very demanding. Therefore, proposed approaches for quick evaluation of scenario generation methods by the quality of decisions they provide are needed to decrease the computational burden and time requirements of scenario evaluation.

1.4 Thesis Structure and Overview

This thesis is structured as follows: it consists of three main chapters, preceded by this general introduction and followed by a general conclusion. The references for each individual section are listed at the end of the corresponding chapters. Each of those main chapters is a journal article manuscript, with the first published in *Wind Energy*, the second under review, and the third in preparation for submission.

The second chapter, titled "Statistical metrics for assessing the quality of wind power scenarios for stochastic unit commitment," explores the properties of a high quality scenario set and the use of statistical evaluation metrics to measure properties of scenarios that are expected to lead to a lower cost in SUC. We borrow some concepts from meteorology; specifically, ensemble forecast verification, and modify them to be used for probabilistic scenarios. Moreover, a new reliability assessment tool for scenarios that may have unequal probabilities is developed. Reliability is the statistical consistency between the probabilistic scenarios and observations. Wind power scenarios are generated by two distinct approaches: a functional approximation approach and a statistical quantile regression approach. Comparative assessment of these two very different scenario generation methods is provided.

The third chapter, titled "Reliability of wind power scenarios and stochastic unit commitment cost," focuses on examining the relationships between the wind power scenario evaluation methods and the SUC simulation results. A study is conducted to solve the SUC and dispatch problem over a historical time period. The statistical approach to eliminate the wind power scenarios that might lead to high costs and to compare the remaining scenario sets is explained. Moreover, the limitations of statistical metrics are explored. It is shown that the wind power scenario generation method and variant that performs the best in SUC could be predicted by the statistical metrics.

The fourth chapter, titled “Scenario generation reliability assessment for two-stage stochastic programs,” focuses on a general approach to assess the scenarios for two-stage stochastic programming problems. We search for approaches to assess scenario quality without having to explicitly solve the stochastic program, which may become intractable due to the number of scenarios and/or size of deterministic (single-scenario) subproblems. We propose approaches for assessing the quality of a scenario generation method using observations of the actual data over a historical period. This approach evaluates the impacts of the scenarios on the costs of single-scenario sub-problems of the associated stochastic program when the first stage variables are fixed to common values. The method is demonstrated by simulation studies and case studies on stochastic server location and stochastic unit commitment problems.

Finally, in Chapter 5 we conclude the thesis with a summary of the results, contributions, limitations, and suggestions for further research.

REFERENCES

[1] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*. New York: Springer, 1997.

[2] S. Takriti, J. R. Birge, E. Long. A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems*, Vol. 11, No. 3, August 1996.

[3] A. Tuohy, P. Meibom, E. Denny, M. O’Malley. Unit commitment for systems with significant wind penetration. *IEEE Transactions on Power Systems*, Vol. 24, No. 2, May 2009.

[4] A. Papavasiliou, S. S. Oren, R. P. O’Neill. Reserve requirements for wind power integration: A scenario - based stochastic programming framework. *IEEE Transactions on Power Systems*, Vol, 26, No. 4, November 2011.

CHAPTER 2: STATICTICAL METRICS FOR ASSESSING THE QUALITY OF WIND POWER SCENARIOS FOR STOCHASTIC UNIT COMMITMENT

A paper published in *Wind Energy*

Didem Sari, Youngrok Lee, Sarah M. Ryan and David Woodruff

Abstract

In power systems with high penetration of wind generation, probabilistic scenarios are generated for use in stochastic formulations of day-ahead unit commitment problems. To minimize the expected cost, the wind power scenarios should accurately represent the stochastic process for available wind power. We employ some statistical evaluation metrics to assess whether the scenario set possesses desirable properties that are expected to lead to a lower cost in stochastic unit commitment. A new mass transportation distance (MTD) rank histogram is developed for assessing the reliability of unequally likely scenarios. Energy scores, rank histograms, and Brier scores are applied to alternative sets of scenarios that are generated by two very different methods. The MTD rank histogram is best able to distinguish between sets of scenarios that are more or less calibrated according to their bias, variability and autocorrelation.

Keywords: Reliability, Energy score, Mass transportation distance, Rank histogram, Brier score.

2.1. Introduction

The wind energy industry is one of the fastest growing renewable energy industries in the world and global wind power capacity continues to grow rapidly. High penetration of wind power requires more sophistication in operational planning to accommodate variability. One of the most significant short-term planning problems for electrical power generation is unit commitment, in which an optimal on-off schedule is found for each thermal generating unit over a given period of time [1]. Unit commitment problems traditionally have been solved by imposing fixed reserve limits to manage uncertainty in load and small amounts of variable generation. However, in systems with a large amount of wind power, cost savings have been demonstrated by solving stochastic unit commitment (SUC) problems with probabilistic scenarios for the wind power trajectory [2, 3, 4, 5].

Errors in the day-ahead wind power forecast and variability in electricity demand create uncertainty in the forecast for net load, which equals the load less the available wind power. A high level of wind penetration increases that uncertainty. In stochastic unit commitment, implicit reserve levels are identified by finding a unit commitment schedule that minimizes expected costs with respect to a set of probabilistic scenarios. The stochastic optimization approach is based on the concept of recourse. In a two-stage model, a single commitment schedule is determined in the first stage by considering how each unit would be dispatched in each scenario of available wind power in the second, or recourse, stage. Compared to the schedule found with fixed reserve limits, costs are saved by committing more resources on days when uncertainty is high and/or expected wind contribution is low, which avoids having to start up additional generating units in real time, and by committing fewer resources on days when uncertainty is low and/or expected wind contribution is high, thus incurring lower start-up and no-load costs. For the stochastic planning approach to be effective, the scenarios must accurately represent the stochastic process for available wind power given information available when the schedule is generated. The scenario time series of wind power amounts should somehow resemble the corresponding observed time series in attributes such as the levels of wind power available at time points throughout the planning horizon, the correlations among these levels, the presence and severity of ramps, etc., and their probabilities should accurately reflect the frequency of similar occurrences.

Recently, considerable effort has been devoted to developing methods for generating wind power scenarios. Different methods yield sets of scenarios that differ quantitatively and qualitatively, in both obvious and subtle ways. A convincing way to evaluate a scenario generation method is to simulate employing the resulting scenarios in stochastic unit commitment while measuring the costs incurred [6]. However, although computational methods for solving stochastic unit commitment problems have improved significantly [4], a simulation study sufficiently thorough to accurately detect meaningful differences among scenario sets is computationally very demanding. Therefore, in this paper we explore the use of statistical metrics to measure properties of scenarios supposed to be desirable for achieving cost savings in stochastic unit commitment. We apply them to distinguish between two very different scenarios generated by approaches: a statistical approach that combines quantile regression with a Gaussian copula [7] and epi-spline approximation approach. Epi-splines and their applications are discussed in [8] and a similar scenario generation method is used for electricity demand in [5]. Because the latter approach yields scenarios that are not necessarily equally likely, we develop and test a new mass transportation distance (MTD) rank histogram to assess whether scenarios with unequal probabilities have similar temporal patterns as the corresponding observations. We present simulation studies to determine MTD rank histogram features that indicate reliability. The histogram evaluation is combined with comparisons of energy scores [9,10] and event-based Brier scores [9] to compare and contrast multiple attributes of the scenario sets.

This paper examines several verification tools that are used to test wind power scenarios for reliability, sharpness, skill and their ability to capture critical characteristics of stochastic processes. We employ energy scores to inform on the forecast skill of scenarios for individual lead times as done in [9]. The energy score has also been used for probabilistic forecasts of surface wind vectors in [10]. However, when applied to multivariate probabilistic forecasts, it has limited ability to discriminate among sets of forecasts with different levels of autocorrelation [11]. Minimum spanning tree rank histograms are used to check reliability of equally likely scenarios [9] or ensemble forecasts [10,12,13]. Smith suggested the use of MST lengths as a scalar pre-rank function for multidimensional forecasts [14]. The MST rank histogram was further studied by Wilks [12] and Gombos et al. [13]. Because all of the scenario sets generated in [9] were equally likely, MST rank histograms were employed to check the temporal dependence structure. However, one of the scenario generation methods presented in our numerical study produces unequally likely scenarios. To incorporate the probabilities, we employ mass transportation distance [15,16] as a pre-ranking function. The mass transportation distance is motivated by stability analysis for use in scenario reduction for stochastic programming [17]. Finally, event-based verification assesses the ability of wind power scenarios to accurately represent ramp up and ramp down events, which can have a large impact on unit commitment and subsequent dispatch costs. Brier scores [18] are applied as an event-based verification approach as in [9].

There has not been much rigorous evaluation of scenario generation approaches according to their performance in stochastic unit commitment. The study reported in [6] is one exception, where the advantages of using SUC formulations over deterministic ones and the importance of probabilistic wind power scenarios are also emphasized. A small study comparing epi-spline load scenarios with Monte Carlo scenario paths in SUC is reported in [19]. Within scenario generation approaches, different variants and parameter settings can produce different sets of scenarios, and simulating their performance in SUC may be computationally prohibitive. The contribution of this paper is to summarize statistical metrics' capabilities and illustrate their potential use as prescreening tools for either equally or unequally likely scenario sets.

The paper proceeds as follows: The existing statistical metrics for scenario evaluation along with our new MTD rank histogram are explained in detail and some simulation studies on MTD rank histograms are provided in Section 2. The wind power scenario generation methods are described in Section 3 including some variations. In Section 4, we compare the results of the two scenario generation methods according to the metrics using wind power forecast and observational data from a U.S. agency. Finally, we conclude in Section 5 with a brief summary and discussion of research directions.

2.2 Verification of scenarios

In this section, some important verification approaches are presented for assessing the quality of scenarios. It is critical to evaluate how well the scenario set reflects the actual wind power output. Some properties of a scenario set are reliability, sharpness, and skill. Reliability refers to the statistical consistency between the probabilistic scenarios and observations [20]. If the relative frequency of occurrence of events assigned a scenario probability tends to be close to the observation, then we accept that scenario set to be reliable or calibrated [21]. Sharpness is the concentration of the scenario distributions. The sharper the scenarios, the less uncertainty they express. What is expected from a forecast is to maximize the sharpness, subject to calibration [20]. Sharpness and calibration are accepted as the components of skill [9]. However, a set of scenarios for stochastic programming has a different purpose than an ensemble of forecasts. A very sharp set of scenarios may not express the uncertainty that decision procedures must consider.

The notation used in this paper is as follows:

$y_d^0 = \{y_{h,d}^0\}$: observed wind power in hour h=1,…,H on day d =1,…, D

$y_d^s = \{y_{h,d}^s\}$: wind power in hour h=1,…,H on day d =1,…, D, in scenario s = 1,…, S

y_d^{0*} : standardized time trajectory, obtained by scaling the wind power levels, y_d^0 according to the installed capacity.

y_d^{s*} : standardized time trajectory, obtained by scaling the wind power levels, y_d^s , according to the installed capacity.

$y_d^{s\circ}$: de-biased wind power on day *d* in scenario *s*

z_d^0 : observed wind power trajectory on day *d* after scaling according to Mahalanobis transformation

z_d^s : wind power trajectory on day *d* in scenario *s* after scaling according to Mahalanobis transformation

p_d^s : probability of occurrence of scenario *s* on day *d*

2.2.1 Energy score

The energy score, a multivariate version of continuous rank probability score [7], has been used to measure the skill of scenarios [6,7]. As mentioned above, skill encompasses both calibration and sharpness. Here we explain the energy score in terms of a distance metric. A statistical distance between two probability distributions *F* and *G* can be defined as [22]:

$$D\left(F,G\right)=2\mathbb{E}\left\|X-Y\right\|-\mathbb{E}\left\|X-X'\right\|-\mathbb{E}\left\|Y-Y'\right\|,$$

where X and X' are independent and identical random vectors having the distribution F , and Y and Y' are independent and identical random vectors having the distribution G . The notation $\mathbb{E}\,||\,.\,||$ represents an expectation of the Euclidean norm. Let $F_d(.)$ be the true probability distribution of wind power generation on day *d* and $\hat{F}_d(.)$ be its estimate. An observation of wind power on day *d*, denoted by y_d^{0*} , is the only available sample point from $F_d(.)$, whereas *S* wind power scenarios $\left\{y_d^{1*},\,.\,.\,.\,,y_d^{s*}\right\}$ can be seen as approximating or having been sampled from $\hat{F}_d(.)$. Then a distance between $F_d(.)$ and $\hat{F}_d(.)$ is computed by;

$$D\left(F_d, \hat{F}_d\right) = 2\sum_{s=1}^{S} p_d^s \left\| y_d^{0*} - y_d^{s*} \right\| - \sum_{s=1}^{S}\sum_{t=1}^{S} p_d^s p_d^t \left\| y_d^{s*} - y_d^{t*} \right\|$$

The energy score is the quantity obtained by dividing $D\left(F_d, \hat{F}_d\right)$ by two:

$$\mathrm{ES}\left(\hat{F}_d, y_d^{0*}\right) = \sum_{s=1}^{S} p_d^s \left\| y_d^{0*} - y_d^{s*} \right\| - \frac{1}{2}\sum_{s=1}^{S}\sum_{t=1}^{S} p_d^s p_d^t \left\| y_d^{s*} - y_d^{t*} \right\|$$

Thus, an energy score can be interpreted as a distance between the true distribution and a scenario distribution of wind power on each day. Therefore, it is a negatively oriented proper score; i.e., lower energy score translates to a higher skill of scenarios [9]. In the case of equally likely scenarios, the formula is simplified as

$$\mathrm{ES}\left(\hat{F}_d, y_d\right) = \frac{1}{S}\sum_{s=1}^{S} \left\| y_d^{0*} - y_d^{s*} \right\| - \frac{1}{2S^2}\sum_{s=1}^{S}\sum_{t=1}^{S} \left\| y_d^{s*} - y_d^{t*} \right\|.$$

A large ES is caused by either the observation being distant from the scenarios or scenarios being too close to each other, or both of these conditions. The ES is not informative with respect to the interdependence structure of the observation or the scenarios [9].

2.2.2 Distance-based rank histograms

The minimum spanning tree rank histogram was developed to verify the reliability of multidimensional ensemble forecasts. Given a set of *m* points connected by edges, a spanning tree is constructed by selecting *m*-1 edges, such that all points are connected. A minimum spanning tree is a spanning tree with the smallest total edge length (Kruskal, 1956). In the context of evaluating scenarios, we find the MST rank by ordering, from smallest to largest, the lengths of the *S*+1 MSTs that are obtained by only scenario points and by successively substituting the observation for each of the scenario points. The rank histogram plots the frequency of the rank among all of the MST lengths of the MST length that is derived from only scenarios. MST rank histogram construction proceeds as follows [14]:

(a) Standardize the set $\left\{y_d^0, y_d^1, \ldots, y_d^s\right\}$ to obtain a standardized observation y_d^{0*}, and standardized scenarios $y_d^{1*}, \ldots, y_d^{s*}$. In the numerical study in Section 4, standardized time trajectories are obtained by scaling the wind power levels according to the installed capacity.

(b) Find the length, l_0, of a MST for the observation from the set $\left\{y_d^{k*} : k \in \{1,...,S\}\right\}$. For each $j = 1,...,S$, compute the MST length, l_j, for scenario *j*, from the set $\left\{y_d^{k*} : k \in \{0,...,S\} \setminus \{j\}\right\}$. In the numerical study when computing the lengths we use the Euclidean (L2) norm; i.e., the distance between y_d^{i*} and y_d^{j*} is $\sqrt{\sum_{h=1}^{H}\left(y_{h,d}^{i*} - y_{h,d}^{j*}\right)^2}$.

(c) Find the MST rank *r*, which is the rank of observation MST length l_0, when $l_0, l_1, \ldots, l_s$ are sorted from smallest to largest. It is an integer between 1 and $S+1$

For an ideally calibrated ensemble of equally likely scenarios, the probabilities of the observation rank falling into any of the bins are equal. Thus, the resulting MST rank histogram should appear uniform. The lowest MST ranks are seen too often for a biased or under-dispersed ensemble, whereas the highest ranks occur too often for an over-dispersed ensemble [10].

MST rank histograms can be used to assess reliability of equally likely scenarios. For scenarios with different probabilities of occurrence, we have devised the mass transportation distance (MTD) rank histogram for the same purpose. In general, the mass transportation distance between two distributions is the minimum cost of transporting the probability from one distribution to the other, where cost is proportional to the distance between supporting points of the distributions [15, 16]. Although in general the MTD is found by solving a linear program, in our application, it is the minimum cost of transporting the probability from the group to the individual. Assuming an edge exists between each pair of points, the trivial solution to the minimization problem uses the tree composed of the edges between each group member and the individual. Thus, the minimum transportation distance from $\left\{y_d^{k*} : k \in \{1,...,S\}\right\}$ to y_d^{0*} can be computed simply as:

$$\sum_{k=1}^{S} \left\| y_d^{k*} - y_d^{0*} \right\| p_d^k .$$

Note that the MTD is identical to the first term in the energy score formula, which measures reliability. This motivates the use of the MTD as a pre-ranking function for reliability assessment. We compute the rank for the observation distance by successively interchanging each scenario, along with its probability, with the observation. In contrast to the MST length as a pre-ranking function, a relatively small distance from scenarios to observation indicates that the observation falls within the convex hull of scenarios; therefore, we order the MTD values from largest to smallest to determine the rank.

Our studies on MTD rank histogram show that it behaves similarly to the MST rank histogram when applied to equally likely scenarios. Its construction is similar to that of the MST rank histogram. It can be constructed by replacing steps *(b)* and *(c)* with steps *(b)'* and *(c)'* as follows:

(b)' Find the MTD for the observation, l_0', which is the distance from the set of scenarios $\left\{y_d^{k*} : k \in \{1,...,S\}\right\}$ to the observation y_d^{0*}. Then compute the MTDs for scenarios, l_j', $j = 1,...,S$, from the set $\left\{y_d^{k*} : k \in \{0,...,S\} \setminus \{j\}\right\}$ to y_d^j. When computing l_j', assign the probability of scenario y_d^{j*}, which is p_d^j, to the observation y_d^{0*}.

(c)' Find the MTD rank *r*, of the observation MTD l_0', when $l_0', l_1', \ldots, l_s'$ are ordered from largest to smallest. It is an integer between 1 and $S+1$.

Fig. 2.1 and Fig. 2.2, respectively, illustrate the constructions of both minimum spanning tree and mass transportation distance lengths that could result from an over-dispersed and under-dispersed ensemble.

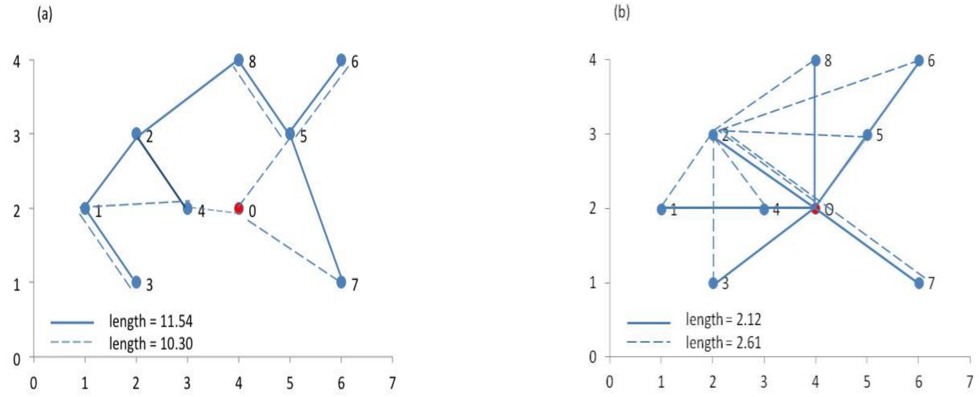


Figure 2.1 A hypothetical example in 2 dimensions is presented for minimum spanning tree and mass transportation distance. S = 8 equally likely scenarios are labeled 1-8 and the corresponding observation is 0. The observation is interior to the scenario points, which causes an over-dispersed ensemble. (a) The solid lines indicate an MST for the scenarios, and the dashed lines indicate an MST that results from the observation being substituted for scenario 2. (b) Similarly, the solid lines indicate the edges used to transport probability from the scenarios to the observation, and the dashed edges are used to transport probability to from all other scenarios plus the observation.

In Fig. 2.1a, substituting the observation for scenario 2 reduces the MST length. The rank of the solid-line MST depends also on the lengths of the other seven MSTs, which are constructed by replacing each of the seven points by the observation in turn. In Fig. 2.1a the lengths of six of them are shorter than 11.54, and only the one that results from replacing scenario 4 by the observation is equal to 11.54. Therefore, the rank of the observation's MST length Fig. 2.1a is 8 or 9 out of 9. In Fig. 2.1b the MTD between scenario 2 and other scenarios along with the observation is 2.61, and the MTD between the scenarios and the observation equals 2.12. Similarly to MST rank, the mass transportation distance rank depends also on the other MTD lengths, which result from replacing each scenario point by the observation in turn. All other MTD lengths are longer than 2.12, which means the rank is 9 out of 9. Thus, the MTD rank agrees with the MST rank in this instance.

We repeat the same process for a biased and/or under-dispersed ensemble in Fig. 2.2.

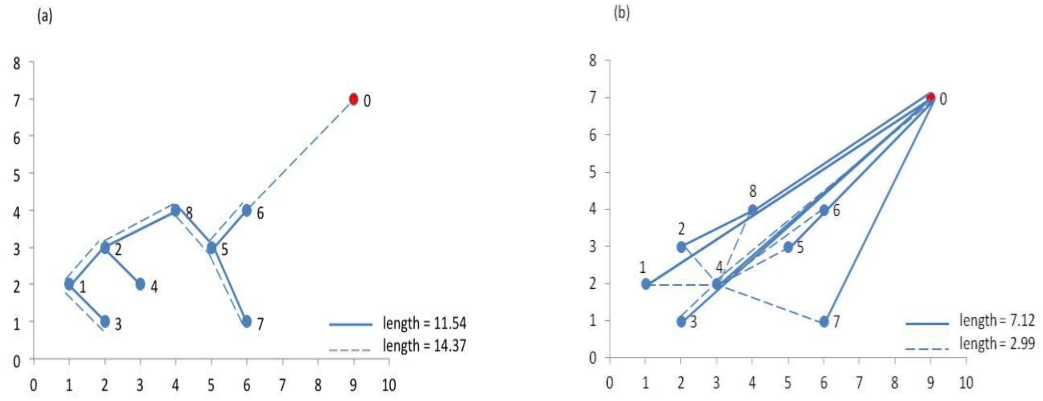


Figure 2.2 The observation 0, is moved to point (9,7) from point (4,2) to obtain an under-dispersed and/or biased ensemble, whereas the 8 scenarios labeled 1-8, have been kept at their same coordinates. The observation becomes exterior to the convex hull of the scenarios. (a) The solid lines indicate an MST for the scenarios labeled 1-8, and the dashed lines indicate an MST that results from the observation being substituted for scenario 4. (b) Similarly, solid edges are used to transport probability from the scenarios to the observation, and dashed edges are used to transport probability to scenario 4 from all other scenarios plus the observation.

In Fig. 2.2a, substituting scenario 4 for the observation yields a smaller rank for that MST length. All of the other seven MSTs also have lengths longer than the solid line’s length. Therefore, the MST rank of the example shown in Fig. 2.2a is 1 out of 9. In Fig. 2.2b the MTD between scenario 4 and other scenarios with the observation, is shorter than the MTD from all scenarios to the observation. The other seven MTDs also have shorter lengths. Because we order the MTDs from largest to smallest, the rank assigned to the observation is 1 out of 9.

In the simulation studies depicted in Figs. 2.3 and 2.4, respectively, MST and MTD rank histograms are constructed for same sets of observations and hourly scenario values, which are all randomly generated from independent normal distributions. For these equally likely simulated scenarios, both types of histograms show the same patterns as distribution parameters are varied. The horizontal axis identifies the bin and the vertical axis measures the frequencies of the ranks that fall into the corresponding bins. Specifically, the MST and MTD rank histogram both display a downward trend for an under-dispersed ensemble ($\sigma_{scen}^2 / \sigma_{obs}^2 < 1$) and an upward trend for an over-dispersed ensemble ($\sigma_{scen}^2 / \sigma_{obs}^2 > 1$), as expected. Flat histograms result when the observation and scenarios are drawn from the same distribution. When the variances are equal, we see a downward trend for larger scenario means, which correspond to bias in the ensemble. Bias over-populates the small ranks similarly as under-dispersion. However, both types of rank histogram appear flat when $\mu_{scen} = 1$ and $\sigma_{scen}^2 / \sigma_{obs}^2 = 2$. This suggests that high variance in the scenarios can compensate for bias [12].

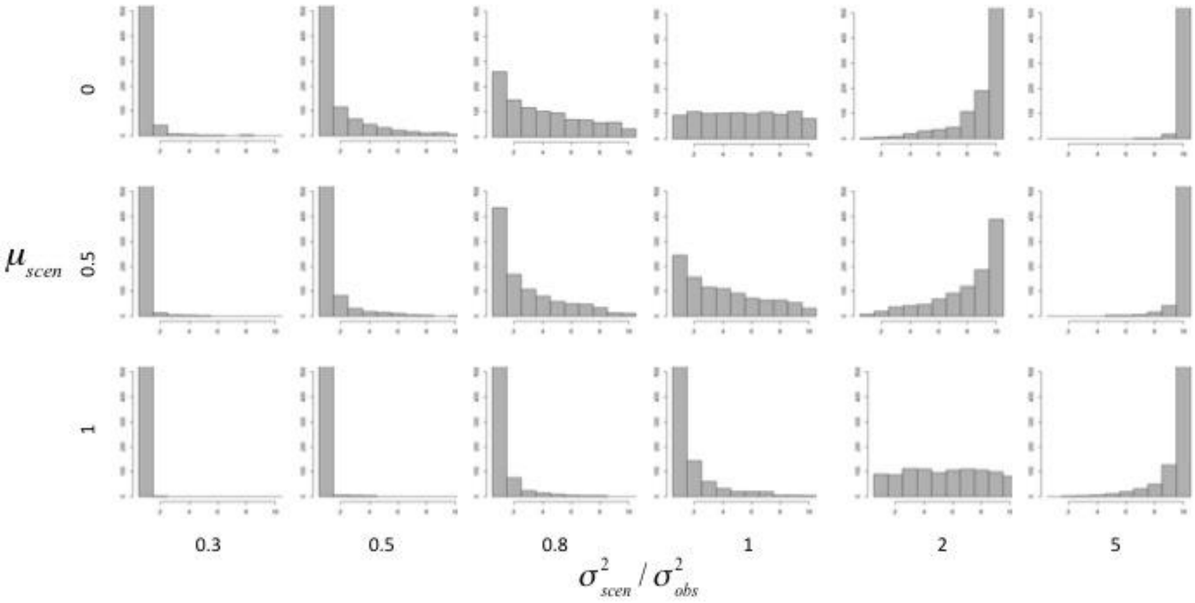


Figure 2.3 Minimum spanning tree rank histogram – simulation study. In this simulation, for each panel, 1,000 ensembles each consisting of one observation and 9 equally likely scenarios, which are vectors of length 24, are sampled. The observation is sampled from a standard normal distribution with mean $\mu_{obs} = 0$ and $\sigma^2_{obs} = 1$. The scenarios are sampled from a normal distribution with mean $\mu_{scen} = 0,\ 0.5,$ or 1. The rows correspond to μ_{scen}, and the columns correspond to the ratio of the scenario variance to observation variance, $\sigma^2_{scen} / \sigma^2_{obs}$.

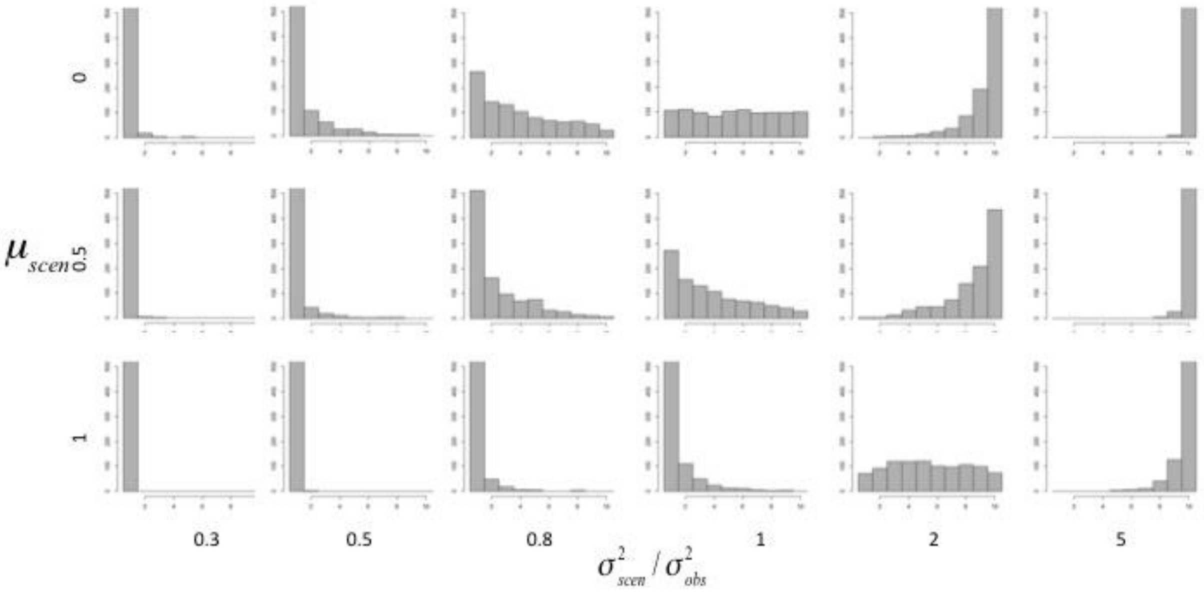


Figure 2.4 Mass transportation distance rank histogram – simulation study with the same setup as in Fig. 3.

In Fig. 2.5 we show the importance of de-biasing for the MTD rank histogram as well as the MST rank histogram. In the left-hand panels, both histograms slope downward because of high bias, even though the scenarios are over-dispersed. To prevent misdiagnosis, Wilks suggested to de-bias the data when constructing MST rank histograms [12]. In the right-hand panels, the data are de-biased according to the following equation:

$$y^{s\circ}_{h,d} = y^{s*}_{h,d} - \frac{1}{D}\sum_{d=1}^{D}\left(\frac{1}{S}\sum_{s=1}^{S} y^{s*}_{h,d} - y^{0*}_{h,d}\right), \text{ for } h = 1, \dots, H \text{ .}$$

The resulting MTD rank histograms appear very similar to the MST rank histograms both before and after de-biasing.

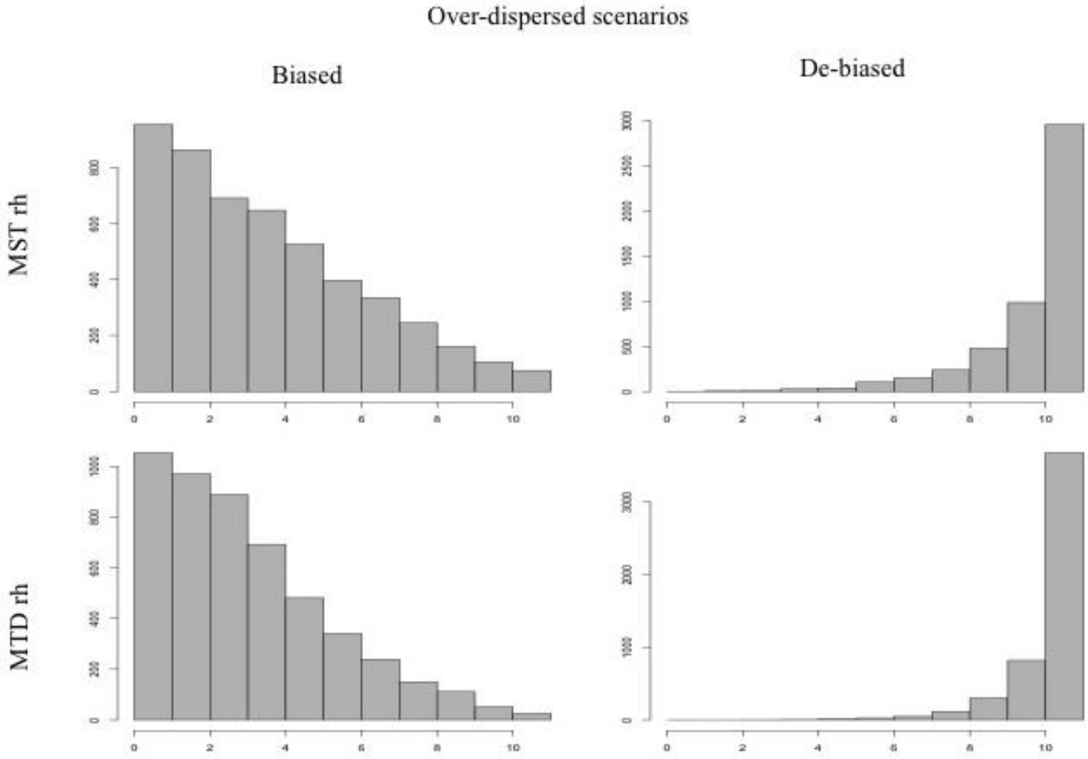


Figure 2.5 MTD and MST rank histograms for over-dispersed scenarios with and without bias - For each panel, 5,000 ensembles each consisting of one observation and 10 scenarios, which are vectors of length 8, are sampled. The observation is sampled from a standard normal distribution with mean $\mu_{obs} = 0$ and $\sigma^2_{obs} = 1$. Scenarios are sampled from a normal distribution with mean $\mu_{scen} = 2.5$ and $\sigma^2_{scen} = 5$.

In the context of wind power, we are particularly interested in assessing whether the autocorrelation, as a way of describing temporal smoothness, of scenarios matches that of observations. In [23] the authors investigate the sensitivity of four different multivariate ranking methods, including minimum spanning tree rank histogram, to miscalibration in the dependence structure. They generate their forecasts from an AR(1) process, whereas their observations follow more complex correlation models. Simulation studies presented in Fig. 2.6 and 7 examine the behaviors of both rank histograms according to autocorrelation. To equalize variances of the marginal distributions the data are scaled according to the Mahalanobis transformation [12].

The Mahalanobis transformation scales the data according to the sample covariance matrix:

$$\mathrm{S}_{scen} = \frac{1}{S}\left[\left(y_d^{0*} - \overline{y}_d^{scen}\right)\left(y_d^{0*} - \overline{y}_d^{scen}\right)^{\mathrm{T}} + \sum_{s=1}^{S}\left(y_d^{s*} - \overline{y}_d^{scen}\right)\left(y_d^{s*} - \overline{y}_d^{scen}\right)^{\mathrm{T}}\right],$$

where

$$\overline{y}_d^{scen} = \frac{1}{S+1}\left(y_d^{0*} + \sum_{s=1}^{S} y_d^{s*}\right)$$

The transformation is a multi-dimensional extension of standardization by subtracting the mean and dividing by the standard deviation:

$$\begin{aligned} z_d^0 &= \mathrm{S}_{scen}^{-1/2}\left(y_d^{0*} - \overline{y}_d^{scen}\right), \\ z_d^s &= \mathrm{S}_{scen}^{-1/2}\left(y_d^{s*} - \overline{y}_d^{scen}\right) \end{aligned}$$

where $\mathrm{S}_{scen}^{-1/2} = D\Lambda^{-1/2}D^{\mathrm{T}}$, D is the matrix whose columns are the eigenvectors of S_{scen}, and $\Lambda^{-1/2}$ is the diagonal matrix containing the reciprocals of the square roots of the corresponding eigenvalues [12].

The MST and the MTD rank histograms behave similarly as the marginal variance and the autocorrelation parameter are varied. For over-dispersed scenarios, as the observation autocorrelation decreases, the histogram becomes flatter; however, an upward trend can still be observed. For under-dispersed scenarios, a downward trend is observed for all levels of autocorrelation levels of the observation but it is less pronounced when the observation autocorrelation is high. If the scenarios and observation have the same autocorrelation and marginal variance, the MST and MTD rank histograms both appear to be flat, as we observe in the middle panels of Figs. 2.6 and 2.7. When the marginal variances of scenarios and observation are the same, the difference between autocorrelations will affect the pattern of both rank histograms. For $\rho_{obs} < \rho_{scen}$, a sloping downward trend and for $\rho_{obs} > \rho_{scen}$ a sloping upward trend are observed in Figs. 2.6 and 2.7.

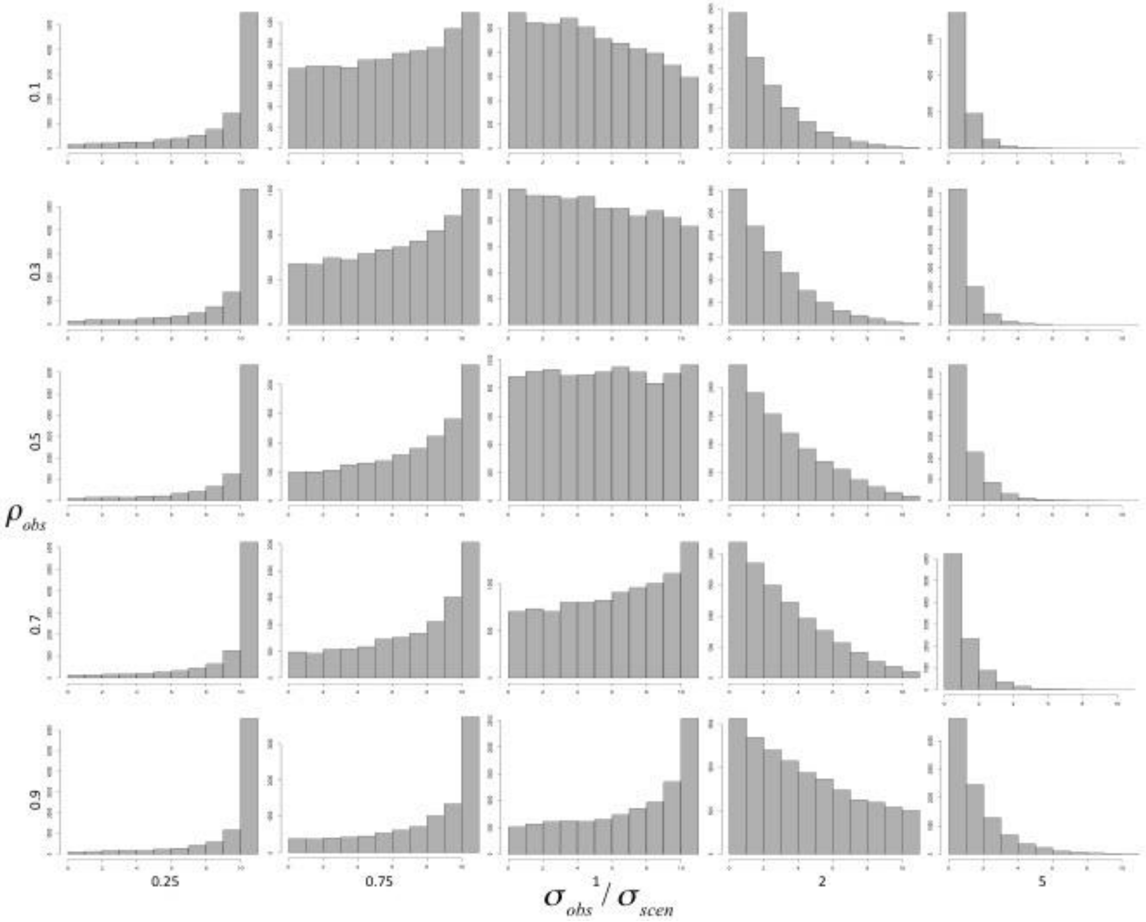


Figure 2.6 Minimum spanning tree rank histogram – simulation study for testing scenarios according to their autocorrelations. In this simulation, for each panel, 10,000 ensembles each consisting of one observation and 10 scenarios, which are vectors of length 8, are sampled. The scenarios are sampled from an AR(1) model, defined as $X_k = \rho X_{k-1} + \varepsilon_k$ with coefficient $\rho_{scen} = 0.5.$ The standard deviation of the marginal distribution of the scenarios is maintained as $\sigma_{scen} = 1$ by adjusting the standard deviation of ε_k. The rows correspond to the ρ coefficients of the observation which is also sampled from the AR(1) model. The columns correspond to the ratios (observation to scenarios) of the standard deviations of the marginal distributions.

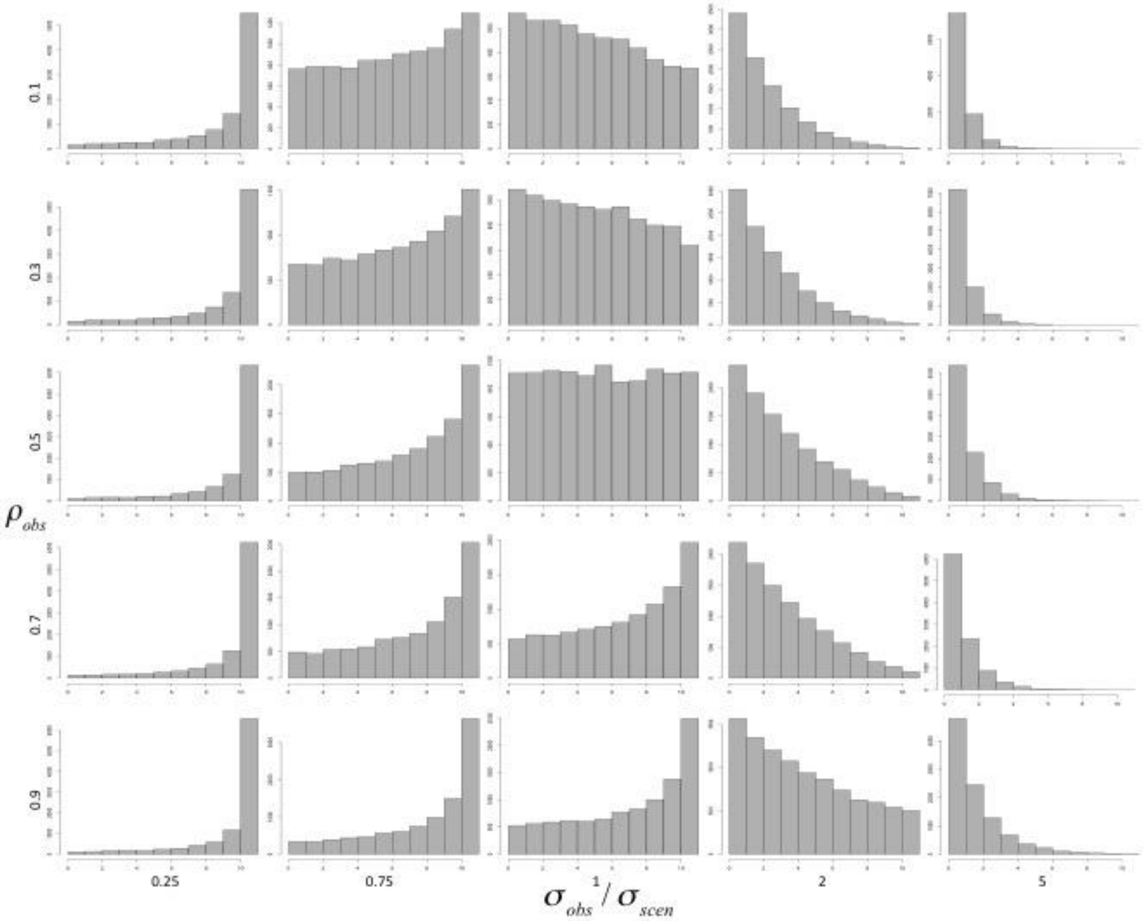


Figure 2.7 Mass transportation distance rank histogram – simulation study for testing scenarios according to their autocorrelations with the same setup as in Fig. 2.6.

Fig. 2.8 further illustrates the patterns of the MTD rank histogram for the case where variances of the marginal distributions of scenarios and observation are equal. The MST rank histograms are flat when the autocorrelations of observation and scenarios are equal. Above the main diagonal where $\rho_{obs} > \rho_{scen}$, they show an upward-sloping trend, which increases with the difference between the autocorrelation levels. When $\rho_{scen} = 0.1$ and $\rho_{obs} = 0.5$, a U-shaped rank histogram is observed. As both ρ_{obs} and ρ_{scen} are increased, an upward-sloping trend appears. For the case where $\rho_{obs} < \rho_{scen}$, below the diagonal, the rank histograms always slope downward but they are flatter when the difference between autocorrelation coefficients of scenarios and observation is smaller.

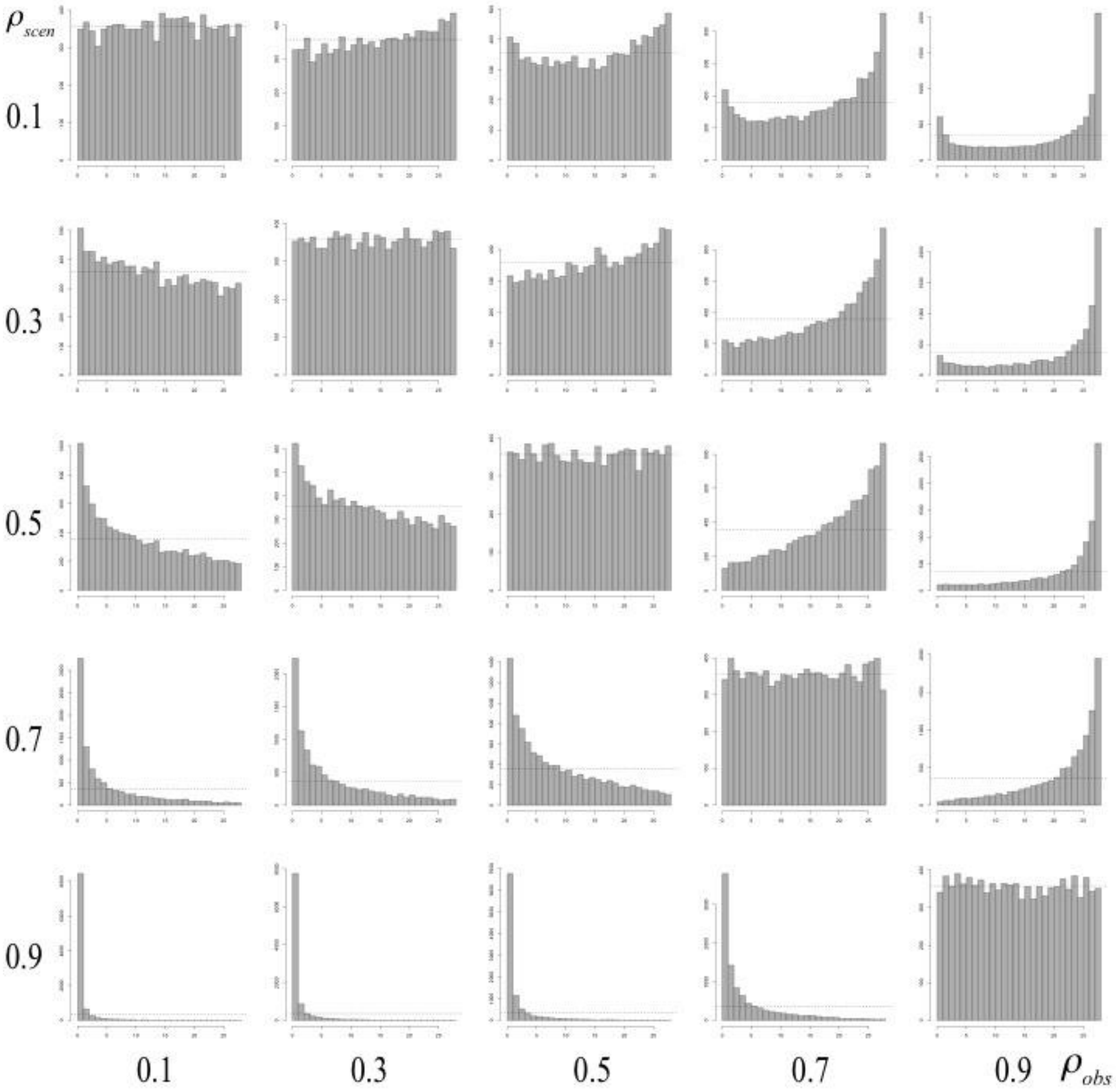
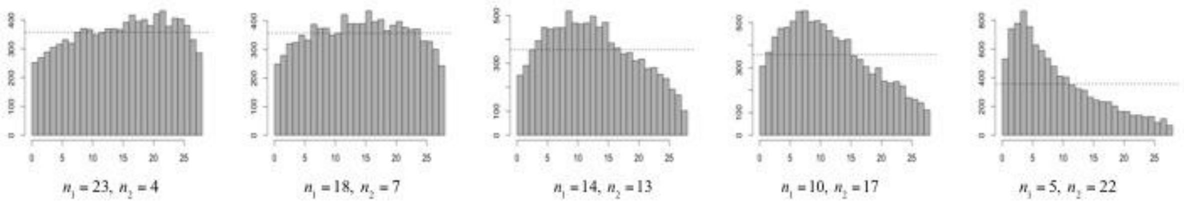


Figure 2.8 MTD rank histograms for $\sigma_{obs}/\sigma_{scen} = 1$ and various combinations of ρ_{obs} and ρ_{scen}

If we generate scenarios with heterogeneous autocorrelation levels, we observe a hill-shaped MTD rank histogram as in Fig. 2.9. This occurs because the presence of both much more and much less smooth scenarios than the observation makes the range of mass transportation distances among scenarios larger. The MTD from the scenarios to the observation will fall in the middle frequently. Overpopulation of the middle ranks results in a hill-shaped MTD rank histogram that is skewed according to the proportions of scenarios with high and low autocorrelation.



$n_1 = 23,\ n_2 = 4$ $n_1 = 18,\ n_2 = 7$ $n_1 = 14,\ n_2 = 13$ $n_1 = 10,\ n_2 = 17$ $n_1 = 5,\ n_2 = 22$

Figure 2.9 MTD rank histograms when $\sigma_{obs}/\sigma_{scen} = 1$ and scenarios with both $\rho_{scen} > \rho_{obs}$ and $\rho_{scen} < \rho_{obs}$ are present. n_1 = the number of scenarios that have AR(1) coefficient $\rho = .1$, n_2 = the number of scenarios that have AR(1) coefficient $\rho = .8$. The observation has an AR(1) coefficient $\rho = .5$.

Certain combinations of over-dispersion and weak correlation can result in a deceptively flat histogram. This is a limitation of both MTD and MST rank histograms. For example, Fig. 2.10 shows relatively flat MTD and MST rank histograms that result from the same setup as in Figs. 2.6-2.9 when $s^2_{scen} / s^2_{obs} = 1.5$ and $r_{scen} = 0.5$, $r_{obs} = 0.1$. The rank histograms could cause the scenarios to be misinterpreted as reliable despite their over-dispersion and higher autocorrelation.

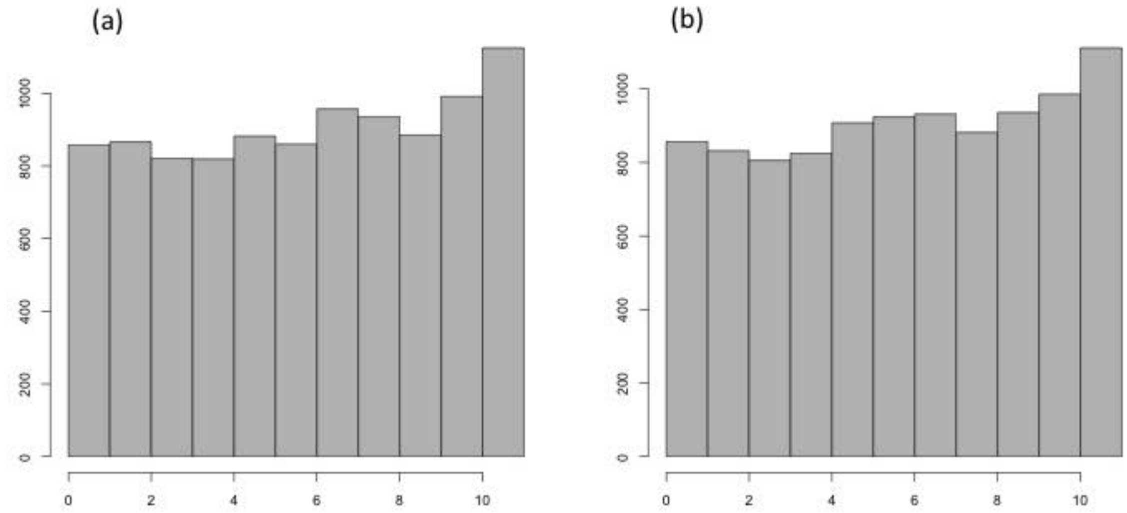


Figure 2.10 (a) MTD and (b) MST rank histograms when $s_{scen}^{2} / s_{obs}^{2} = 1.5$ and $r_{scen} = 0.5$, $r_{obs} = 0.1$

In summary, the shape of the MTD rank histogram closely corresponds to that of the MST rank histogram when applied to equally likely scenarios. It can also be used to diagnose higher, lower, and mixed levels of autocorrelation in the scenarios compared to the observation. To verify its use with unequally likely scenarios, we repeated the same study of the MTD rank histogram as in Fig. 2.7 with the added step of randomly (without replacement) assigning a probability drawn from the set $\left\{ 2i / \left(S \left(S+1 \right) \right), i = 1, \ldots, S \right\}$ to each of the S scenarios generated. The MTD rank histograms showed the same patterns with varying parameters as in Fig. 2.7.

2.2.3 Event-based verification

Event-based verification can be used to explore the scenarios' ability to represent some specific characteristics of stochastic processes as done in [9]. For this verification type, first, it should be determined which stochastic process characteristics are critical to capture. The events can then be defined to detect these critical characteristics.

For instance, a significant gradient event is defined in [9], which is the "maximum absolute variation being greater than a determined threshold in a determined finite duration beginning at a time point". The event parameters are the threshold ξ, and the duration κ. By changing the parameters ξ and κ, different specific events can be defined. Similarly to the significant gradient event, we define ramp up and ramp down events as the "maximum increase and maximum decrease being greater than or equal to ξ, in κ hours beginning at time point h respectively". For wind power scenarios, we are particularly interested in ramp down events because an unexpected loss of a significant amount of wind power could trigger the need for expensive peaking generators to be brought into service. In Section 4, we tested wind power scenarios according to both ramp down and ramp up events.

An indicator variable, denoted as $1\{.\}$, takes value 1 if the event occurs or 0 otherwise. Ramp events are defined as follows for a given time series:

$$\text{RampUp}(y_d; h, \kappa, \xi) = 1\left\{\exists\ i \in \{0,1,...,\kappa-1\} \quad \text{s.t.} \quad y_{(h+\kappa),d} - y_{(h+i),d} \geq \xi\right\}$$

$$\text{RampDown}(y_d; h, \kappa, \xi) = 1\left\{\exists\ i \in \{0,1,...,\kappa-1\} \quad \text{s.t.} \quad y_{(h+i),d} - y_{(h+\kappa),d} \geq \xi\right\}$$

Denoting the parameter set as $\theta = \left(h, \kappa, \xi\right)$, $\mathrm{RampUp}(y_d^0;\theta)$ and $\mathrm{RampDown}(y_d^0;\theta)$ define the ramp up and ramp down events for observed time series on day *d* beginning at time *h* within a time window of length κ. For the scenarios, the event probabilities can be defined mathematically as:

$$\mathrm{P}_{h,d}[\mathrm{RampUp}(y_d^s;\theta)] = \sum_{s=1}^{S} \mathrm{RampUp}(y_d^s;\theta) p_d^s$$

$$\mathrm{P}_{h,d}[\mathrm{RampDown}(y_d^s;\theta)] = \sum_{s=1}^{S} \mathrm{RampDown}(y_d^s;\theta) p_d^s$$

The probability-weighted average of indicator variables for the scenarios takes a value in the interval [0,1]. The Brier score is a strictly proper score to assess these binary situations, which depend on the occurrence and non-occurrence of the event, as applied in [9]. The Brier score is the sum of squared distances between the observation indicator and scenario average [18]. A daily Brier score can be computed as:

$$\mathrm{Bs}(d)_{daily} = \frac{1}{(H-\kappa)} \sum_{h=1}^{H-\kappa} \left(\mathrm{P}_{h,d}[\mathrm{RampDown}(y_d^s;\theta)] - \mathrm{RampDown}(y_d^0;\theta) \right)^2 \quad \text{for } d = 1,\ldots,D$$

In Section 4, we also examine the frequencies of hourly Brier scores:

$$\mathrm{Bs}(h,d)_{hourly} = \left(\mathrm{P}_{h,d}[\mathrm{RampDown}(y_d^s;\theta)] - \mathrm{RampDown}(y_d^0;\theta) \right)^2 \quad \text{for } h = 1,\ldots,H-\kappa, \quad d = 1,\ldots,D.$$

Brier scores measure the degree of correspondence between scenarios and observation based on the event occurrence. Brier scores are lower for scenarios that accurately reflect the event's occurrence.

2.3 Wind power scenario generation methods

We use the methods described above to compare the results of two distinct methods for generating scenarios of short-term wind power generation. Given a forecast time series for amounts of wind power available on the next day, the major challenges in generating scenarios (i.e., alternative time series, each with a probability attached) include modeling the marginal distributions of forecast error at each time point, considering dependencies among these marginal distributions, and building sequences of wind power values that respect the distributions and temporal dependencies. The two general approaches for wind power scenario generation considered in this paper are compared according to these aspects in Table 2.1. Note that the quantile regression approach yields equally likely *sample points* from the estimated joint distribution while the epi-spline approximation approach results in a collection of time series that, together with their probabilities, *approximate* the stochastic process for wind power. In the following two subsections the approaches are described in more detail to explain some of their variants considered in the numerical study.

Table 2.1 Overview of scenario generation approaches

Approach	Epi-spline approximation with information [19]	Quantile regression with copula [7]
Marginal distribution for each time point	Epi-spline approximation of log of error density based on historical errors within a forecast cluster	Linear interpolation of quantiles of forecast error estimated by quantile regression
Intertemporal dependence	Conditional distributions of forecast errors based on categorizations of forecast at certain time points	Gaussian copula applied to marginal distributions to approximate joint distribution
Scenario construction	Conditional expected values within segments of conditional forecast distributions	Monte Carlo samples from joint distribution of forecast errors added to given forecast

One-day wind power output scenarios were generated based on day-ahead wind power forecast data. We followed a “leave-one-out” methodology when generating short term wind power scenarios by both methods. For scenarios generated on day *d*-1 for day *d*, the training set consisted of the whole data range except day *d*, whereas the test day was day *d*.

2.3.1 Wind power scenario generation by quantile regression with Gaussian copula approach

The actual wind power generated at hour *h* on day *d*, $y_{h,d}$, can be observed immediately at the end of hour *h* on day *d*. On day *d*-1 we obtain a vector of day-ahead wind power forecasts (DWPF)

$$\hat{y}_d = \left(\hat{y}_{1,d}, \hat{y}_{2,d}, \ldots, \hat{y}_{24,d}\right).$$

Thus, a day-ahead wind power forecast error (DWPFE) can be observed at the end of each hour *h* on day *d*:

$$e_d = \left(e_{1,d}, e_{2,d}, \ldots, e_{24,d}\right),$$

where

$$e_{h,d} = y_{h,d} - \hat{y}_{h,d}.$$

In this method, actual wind power output, DWPF, and DWPFE are all assumed to be normalized by wind power capacity and denoted as $y_d^{0*}, \hat{y}_d^*, e_d^*$, respectively, so that $\left(y_{h,d}^{0*}, \hat{y}_{h,d}^*\right) \in [0,1]^2$ and $e_{h,d}^* \in \left[0 - \hat{y}_{h,d}^*, 1 - \hat{y}_{h,d}^*\right]$. On day *d*, after DWPF $\hat{y}_d^*$ is obtained, we estimate a distribution of DWPFE

$$F_{h,d}\left(e \middle| \hat{y}_d^*\right) = P\left(e_{h,d}^* \le e \mid \hat{y}_d^*\right)$$

for each hour *h* by linearly interpolating a predicted τ - quantile of $e^{*}_{h,d}$ for each τ in a predefined set of quantiles $\mathbb{T}$ (e.g. $\{.05, .10, \ldots, .95\}$). It is assumed that the 0.00 quantile of the predictive forecast is 0 and the distribution below the 0.05 quantile is modeled as a linear interpolation between the 0 and 0.05 quantiles. Similarly, the 1.00 quantile of the predictive forecast is assumed to be 1 and the distribution above the 0.95 quantile is linearly interpolated. These assumptions may lead to extreme scenarios with unrealistically large differences from the forecast. For better results, the predictive distributions should be parameterized with exponential tails, thus reflecting the unlikeliness of extreme events [7]. Each quantile of $e^{*}_{h,d}$ is predicted by using quantile regression models on

$$O_d = \left\{ \left(\hat{y}^{*}_{1}, e^{*}_{1} \right), \ldots, \left(\hat{y}^{*}_{D}, e^{*}_{D} \right) \right\} \setminus \left\{ \left(\hat{y}^{*}_{d}, e^{*}_{d} \right) \right\}.$$

The development described above is elaborated in [7]. Here, we introduce two variants on constructing predictor variables for quantile regression models. First, we conduct dimension reduction to improve the reliability of the regression models. The original DWPF is highly inter-correlated 24-dimensional data. We define the following five models of transformed DWPF:

- model 1: A single forecast datum for the particular study period *h*: $\hat{y}^{*}_{h,d}$
- model 2: model 1 + forecasts for an hour before and after. This may outperform model1 if there is inaccurate time prediction called phase error [15].

- model 3: Principal components that take account of the major proportion of variances in the DWPF data matrix. In this study, four components explain over 99% of the total variance-covariance in the training data matrix.
- model 4: model 3 + principal components that take account of the major proportion of variances in local differences within DWPF:

$$(\hat{y}^*_{2,d} - \hat{y}^*_{1,d}, \hat{y}^*_{3,d} - \hat{y}^*_{2,d}, \ldots, \hat{y}^*_{24,d} - \hat{y}^*_{23,d}).$$

In this study, five components explain over 90% of total variance-covariance in the local difference data matrix.

- model 5: model 4 on an extended DWPF that includes forecasts for two hours before and after the forecast day:

$$\hat{y}^*_t = (\hat{y}^*_{(-2),d}, \hat{y}^*_{(-1),d}, \ldots, \hat{y}^*_{24,d}, \hat{y}^*_{(+1),d}, \hat{y}^*_{(+2),d}$$

The second variant is to use a spline function to incorporate a possible nonlinear relation between a quantile of forecast error and DWPF. The number of the degrees of freedom (DF) represents the number of basis functions of each regressor, which implies the complexity of nonlinearity between each regressor and the forecast error. We applied natural cubic spline with up to 3 basis functions (DF=1,2,3). By combining five dimension reduction models and three spline functions, we construct 15 different DWPFE distributions for each *h.* A linear interpolation of estimated quantiles may need some exception handling to make sure that $F_{h,d}(.)$ is monotonically increasing and the range of DWPFE is realistic.

After estimating $F_{h,d}\left(.\right)$ for each *h* and *d* we transform the training forecast error $e^{*}_{h,d}$ into normally distributed random variables

$$\mathrm{z}_{h,d} = \Phi^{-1}\left(\hat{F}_{h,d}\left(e^{*}_{h,d} \mid \hat{y}^{*}_{d}\right)\right),$$

where Φ is the cdf of a standard normal distribution.

Let

$$Z\left(d\right) = \begin{bmatrix} z_{1,1} & z_{1,1} & \cdots & z_{24,1} \\ z_{1,2} & z_{2,2} & \cdots & z_{24,2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{1,d-1} & z_{2,d-1} & \cdots & z_{24,d-1} \\ z_{1,d+1} & z_{2,d+1} & \cdots & z_{24,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ z_{1,D} & z_{2,D} & \cdots & z_{24,D} \end{bmatrix} = \begin{bmatrix} z_{1} \\ z_{2} \\ \vdots \\ z_{d-1} \\ z_{d+1} \\ \vdots \\ z_{D} \end{bmatrix}$$

We generate a scenario of transformed DWPFE by

$$z^{s}_{d} \sim N\left(\mu_{0}, \hat{\Sigma}_{d}\right)$$

where μ_{0} is a 24-dimensional zero vector and the variance-covariance matrix

$$\hat{\Sigma}_{d} = \frac{1}{D-1} Z\left(d\right)^{\mathrm{T}} Z\left(d\right).$$

Next, we generate a DWPFE scenario e^{s*}_{d} as $e^{s*}_{d} = F^{-1}_{h,d}\left(z^{s}_{d}\right)$. Finally a scenario of day-ahead wind power output is computed by adding the scenario of DWPFE to DWPF as

$$y^{s*}_{d} = \hat{y}^{*}_{d} + e^{s*}_{d}\ .$$

Each scenario is assumed to occur with probability

$$P_s = \frac{1}{S}, \forall s \in \{1, \cdot \cdot \cdot, S\}.$$

2.3.2 Wind power scenario generation by epi-spline approximation approach

The goal of this functional approximation approach is to sparingly approximate, rather than sample from, the error distributions, while incorporating available information. The three main steps are segmentation, forecast error distribution estimation, and path construction.

The process begins with the DWPF, $\hat{y}_d = \left(\hat{y}_{1,d}, \hat{y}_{2,d}, ..., \hat{y}_{24,d}\right)$, and the observed wind power, $y_d = (y_{1,d}, y_{2,d}, ..., y_{24,d})$, for each day *d* in the training set. Our goal is to characterize the distributions of the forecast performance so that in the future when given a forecast, we can produce probabilistic statements about possible observations.

Using exponential epi-splines [24], for each hour h an error density function is approximated from the forecast errors $\left\{e_{h,d}, d = 1, ..., D\right\}$, and then numerically integrated to obtain an hourly error cdf.

A key concept is *segmentation* of the data so that data from similar conditions are grouped together. The simplest form of segmentation is to group similar amounts of wind power together, as was suggested in [25,26] who cite [27] as the original work.

In operation, we are given a 24-hour wind forecast and asked to provide a distribution of the forecast for certain hours in that forecast. To do that, we find a distribution for *similar hours* in the historic data, which means we use data from the same data segment as the forecast. We segment wind based on two main attributes: the magnitude of the power forecast and the *derivative pattern*.

The magnitude is taken into account by using only those historic hours with a forecast value within a *window*. The width of the window is controlled by a parameter (typically 0.4) that gives the fraction of the distribution centered at the forecast to include in the window; i.e,, approximately the fraction of the observations to include.

The derivative pattern is a bit more involved. For each hour we compute the derivative one hour before, one hour after, and at the hour. Each derivative is classified as small, substantially negative, or substantially positive with the meaning of “substantially” controlled by a parameter. Hence, for any hour there are nine possible patterns. Because some patterns do not have very much historic data, we cluster the patterns using a metric in the space of error distributions to control the clustering. Patterns with similar error distributions are put in the same cluster, which is done in a pre-processing step along with assignment of each historic hour to a cluster. Then, when an error distribution for some hour in the forecast is requested, the derivative pattern for the forecast is determined and only those hours that are in the same cluster and within the magnitude window are used to compute the error distribution.

Paths are constructed in a fashion based on Rios et al. [19]. One difference is that for wind there is no re-forecasting and segmentation is done as described above, not by error category as in [19]. A few hours are selected by the analyst to be *day part separators* (dps).

The probability computations assume that the hours are far enough apart so that correlations in forecast errors between them can be ignored, which has been verified for the studies conducted so far.

At each dps, cutting points on the error distribution are computed and used to compute *skeleton points*, each of which is an expected value conditional on being between a pair of cutting points. The difference in the cutting points is the probability assigned to the skeleton points. A list of skeleton points with one for each dps is called a *skeleton* and, under the assumption that all skeletons are enumerated and that the errors between dps are uncorrelated, the probability attached to a skeleton is simply the product of the probabilities attached to its skeleton points. To connect the points and provide values for the hours between dps the deviation from the forecast is linearly interpolated. This process completes specification of a scenario with serial dependence between the hours based on the forecast dependence and an attached probability.

The number of scenarios generated equals the number of dps multiplied by one less than the number of cutting probabilities. For example, in the numerical study dps consist of hours 0, 12, and 23, to divide the day into two intervals, and 4 probability values including 0 and 1 are used to cut the error distributions into segments. If cutting probabilities (0, 0.1, 0.9, 1) are used then skeleton points computed have probabilities approximately equal to 0.1, 0.8, and 0.1, corresponding to the lower tail, middle, and upper tail of the distribution. (These probabilities are not exact because the conditional distributions are discretized based on the historical data.) Therefore, the probabilities associated with the paths approximately equal $(0.1)^3$ for a scenario inhabiting the tails at all dps, $(0.1)^2(0.8)$ for a path through tails at two dps and the middle of distribution at one dps, $0.1(0.8)^2$ for a path in the tail at one dps and the middle at two dps's, and $(0.8)^3$ for the path that represents the middle of the error distribution at each dps.

2.4 Example application of the verification approaches

We used day-ahead forecast and observational data from the Bonneville Power Administration in a recent year to generate scenarios by both methods. Scenarios were generated by the quantile regression with Gaussian copula approach according to 5 transformation functions (model 1 through 5) and 3 levels of nonlinearity (DF 1 to 3) for the natural B-spline function as explained in Section 3.1. This resulted in 15 quantile regression models, labeled below as QR(*m*,*n*), for model *m* with *n* DF. The epi-spline approximation method was used with two different sets of cutting points $\left(0, p_1, p_2, 1\right)$ and the resulting scenario sets labeled as EPI($p_1, \ p_2$). For each day, 27 scenarios were generated by each method.

2.4.1 The BPA dataset

Bonneville Power Administration (BPA) is a federal non-profit agency based in the Pacific Northwest of the U.S. that markets wholesale electric power. BPA works with wind project owners to develop more accurate long-term and short-term wind forecasts. More information on wind power forecasting methodology by BPA can be found from [28]. Our focus is given to data from 2012-10-01 to 2013-09-30, based on private communication with BPA.

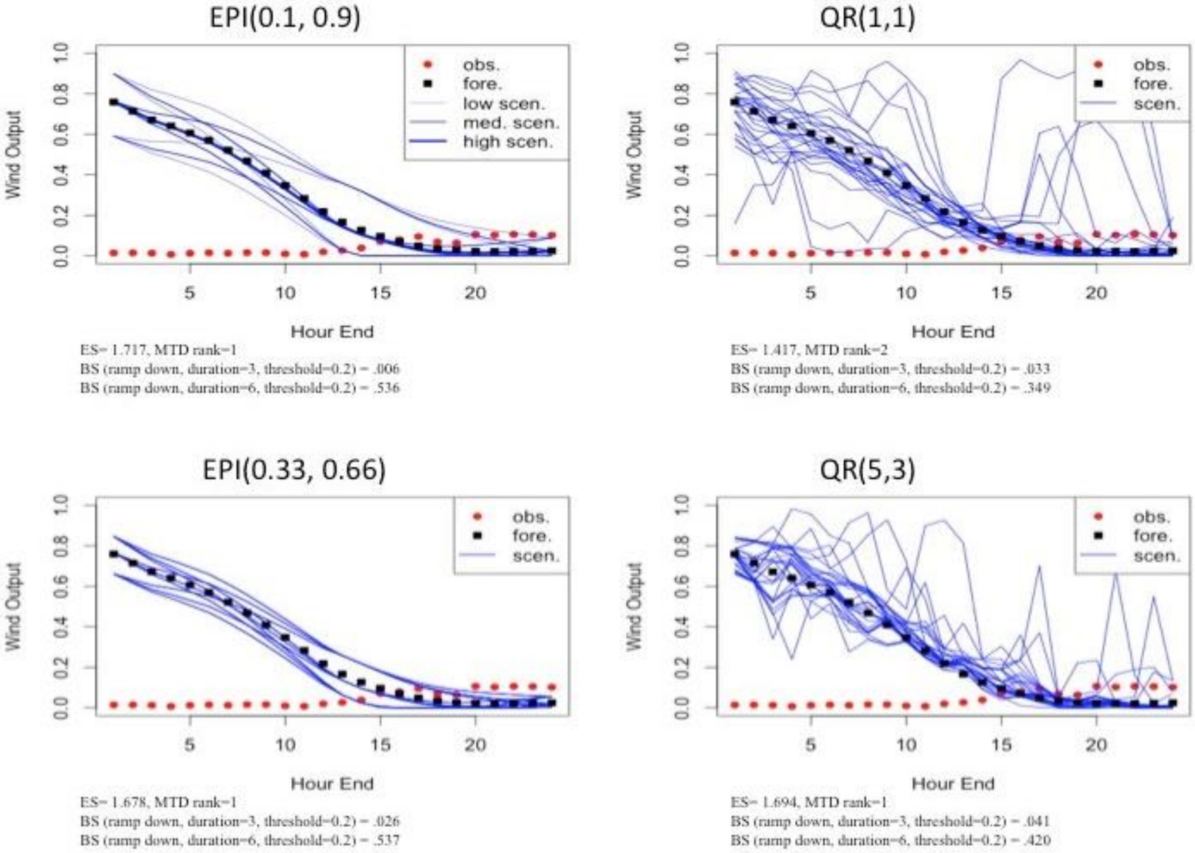
Forecast data were obtained from [29], item number 3; and [30], which provides monthly spreadsheets for “BPA wind power forecasting data”. Each month includes three files; maximum, minimum, and average. We used the “AVG BPA Wind Power Forecast” file, which includes the expected average generation over the hour. Forecasts are identified by UTC stamp and extend over 72 hours. Hr01 represents the first hour of the forecast or next hour. We extracted the forecast generated at 11 a.m. Pacific time, which is 18:00 Greenwich mean time (UTC) during daylight savings time and 19:00 UTC during normal time, on day *d*-1, for the 24 hours of day *d* in columns labeled as Hr13…Hr36. When generating scenarios by the quantile regression method according to model 5 explained in Section 3.1, we need 2 hours of extended forecast data. These forecast values, denoted as $\hat{y}_{(-2),d}, \hat{y}_{(-1),d}, \hat{y}_{(+1),d}, \hat{y}_{(+1),d}$, were obtained from columns Hr11, Hr12, Hr37 and Hr38, respectively.

For the observations, we used the total wind generation value in the first 5-minute interval from item number 5 on [29]. When normalizing the observation and forecast data, we used wind generation capacity available from [31].

A few days within this date range were ignored because of missing information or noisy data. The date when daylight savings time began (2013-03-10) was omitted. Moreover, a few days were omitted as abnormal because they were labeled by BPA as abnormal; specifically, we omitted days that had 4 hourly forecasts with wind states greater than or equal to 2 or less than or equal to -2. The wind states can be found from item 12, “Data for BPA balancing reserves deployed and BPA states” of [29]. In addition to these, days with missing information in either forecast or observation were not included in the scenario generation data. In this one-year period there were a total of 22 disregarded days, leaving a leave-one-out training window length of D=343 days.

2.4.2 Verification of BPA scenarios

In Fig. 2.12-2.14, we show the observed wind power and the scenarios generated by different variations of the two approaches on selected days along with those days' Energy scores (ES), MTD ranks and Brier scores. Fig. 2.12 illustrates the effects of a bad forecast, and Fig. 2.13 shows the results for a day when the wind output and forecast are both very low. These are the most extreme days in our dataset. We distinguish the scenarios generated by the EPI(0.1, 0.9) according to their approximate probabilities and label them as high, medium, or low (probability). Because the probabilities for scenarios generated with EPI(0.33, 0.66) are very similar, we did not distinguish their probabilities in the plots. As mentioned in Section 3, the quantile regression method generates equally likely scenarios. On 2012-11-08, as shown in Fig. 2.12, the scenarios are very far from the observation, which results in very high ES for all of the scenario sets. This is evidently because of the bad wind power forecast. Although the first term in the ES is very large, the dispersion of scenarios can reduce the score. For example, the scenarios generated by the QR(1,1) are more dispersed than the other scenarios sets and have a lower ES. The observation is exterior to the scenarios due to under-dispersion and/or high bias, and this condition is detected by the low MTD ranks for all scenario sets. Compared to the other days, Brier scores are relatively high as expected. Conversely, on 2013-03-26 shown in Fig. 2.13, the scenarios are very close to the observation; thus, the energy scores are very low. Because the first term of the ES is small for all scenario sets, the sharper scenarios generated by EPI(0.33, 0.66) achieve a lower ES. This scenario set is very close to the observation and sharp for that particular day. Brier scores are near or equal to zero for all models. The quantile regression scenarios give higher MTD ranks than the epi-spline scenarios because their wide range causes the observation to lie in their interior. Fig. 2.14 represents a more typical day. The EPI (0.33, 0.66) scenarios have a low ES but do not contain the observed wind power. The quantile regression scenarios envelop the observation but exhibit much higher volatility than either the forecast or the observed wind power levels.



ES= 1.717, MTD rank=1
BS (ramp down, duration=3, threshold=0.2) = .006
BS (ramp down, duration=6, threshold=0.2) = .536

ES= 1.417, MTD rank=2
BS (ramp down, duration=3, threshold=0.2) = .033
BS (ramp down, duration=6, threshold=0.2) = .349

ES= 1.678, MTD rank=1
BS (ramp down, duration=3, threshold=0.2) = .026
BS (ramp down, duration=6, threshold=0.2) = .537

ES= 1.694, MTD rank=1
BS (ramp down, duration=3, threshold=0.2) = .041
BS (ramp down, duration=6, threshold=0.2) = .420

Figure 2.11 Wind power scenarios generated for day 2012-11-08

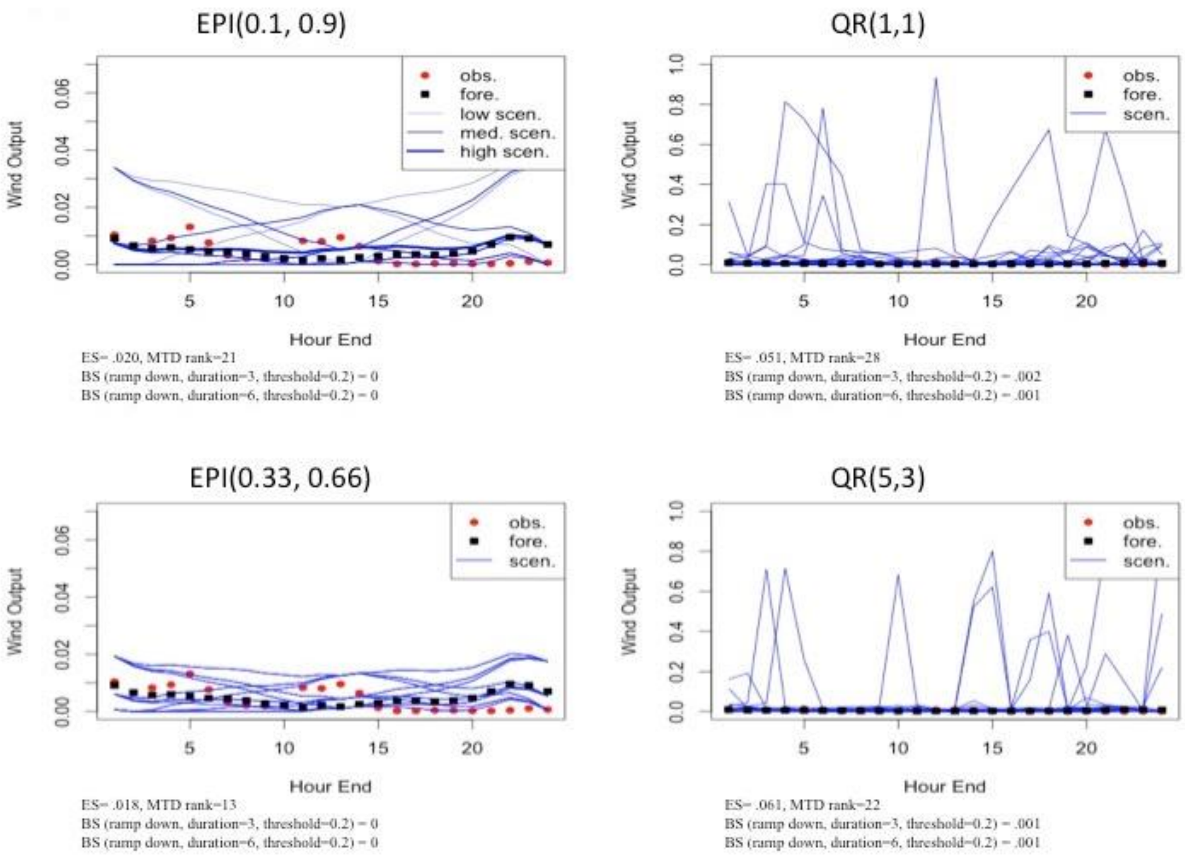


Figure 2.12 Wind power scenarios generated for day 2013-03-26. Note the difference in scale between the left- and right-hand panels.

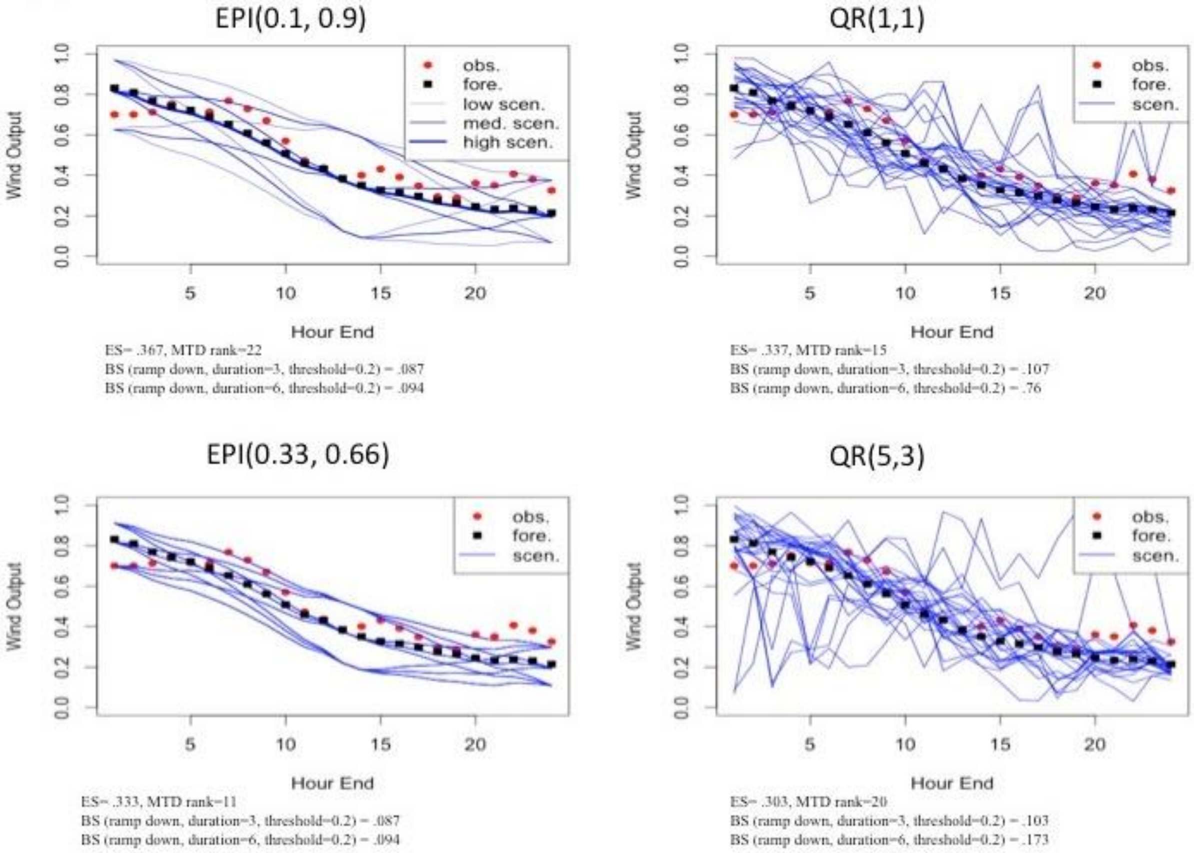


Figure 2.13 Wind power scenarios generated for day 2013-04-08

The means and standard deviations of energy scores for all scenario sets are provided in Tables 2.2 and 2.3. In general, they are very similar and fairly low. Because the differences in score means between the different methods are small compared to the standard deviations, it is hard to draw any conclusion. We applied the pair-wise statistical tests for equal performance [32] to see if there is a significant difference among the scenario sets. According to paired t-tests there were only a few significant differences. The QR(4,3) scenarios had a higher mean ES than all of the other quantile regression scenario sets. Also, EPI(0.1, 0.9) had higher mean ES than EPI(0.33, 0.66) and all of the quantile regression scenario sets except

QR(4,2), QR(4,3), and QR(5,3). The mean ES of the EPI(0.33, 0.66) scenarios was lower in the pairwise comparison than any of the other scenario sets. Thus, according to the energy score, EPI(0.33, 0.66) has the most skill.

Table 2.2 Means and standard deviations of energy scores for scenarios generated by quantile regression with Gaussian copula approach with various combinations of models and nonlinearity

	Model 1	Model 2	Model 3	Model 4	Model 5
DF=1	0.304528 (.180)	0.304323 (.179)	0.303317 (.181)	.303545 (.184)	.305380 (.182)
DF=2	0.303961 (.179)	0.304481 (.185)	0.307880 (.183)	.309423 (.186)	.307906 (.185)
DF=3	0.303177 (.184)	0.305345 (.182)	0.307104 (.184)	.316676 (.188)	.309297 (.182)

Table 2.3 Means and standard deviations of energy scores for epi-spline approximation approach scenarios with different cutting probabilities

	ES (std. dev. of ES)
Shape: (0-0.1-0.9-1)	0.31621 (0.209)
Shape: (0-0.33-0.66-1)	0.29373 (0.195)

We conjecture that for problems such as stochastic unit commitment, the reliability of scenarios is more important than their sharpness. If the actual wind power level exceeds the highest level among scenarios, an opportunity cost would be incurred by having committed too many thermal generators on the day ahead. Actual wind power falling below all the scenarios would likely necessitate the dispatch of expensive peaking units. Thus, it could be risky to only depend on the ES, which encompasses both reliability and sharpness. A low ES that is obtained due to sharp scenarios (which are not perfectly reliable) could misleadingly indicate high quality of scenarios, which do not actually represent the uncertainty properly. The MTD rank histogram better identifies whether the scenarios having lower ES are well calibrated.

MTD rank histograms after de-biasing and scaling are displayed in Fig. 2.14. The Mahalanobis transformation scales the data according to the sample covariance. The sample covariance matrices computed for the epi-spline scenarios were not actually positive definite as shown by slightly negative eigenvalues. To remedy this and allow the computation of the required square roots, we employed a method to find the nearest positive definite covariance matrix [33]. The Cramèr-von Mises statistics for all four rank histograms indicate rejection of a hypothesis of uniformity at the 1% significance level. However, the statistic for the rank histogram in (d) is very close to the critical value of 0.33. The smallest rank is overpopulated in MTD rank histogram (b) for scenarios generated by EPI(0.33, 0.66), which means the temporal dependence structure in the observation is overestimated. The one in (a) generated by EPI(0.1, 0.9) is relatively flat. In (c) we can observe a hill shape, which indicates that scenarios with both lower and higher levels of autocorrelation than the observation are generated by QR(1,1). The histogram generated by QR(4,3) in (d) is flatter than (c) and, overall, indicates the best match of autocorrelation and variance levels between scenarios and observations. The hill shapes observed in (c) and (d) might be attributed to the linear interpolation of the marginal tails, which caused excursion of scenarios far from the pack. This increased the number of scenarios with lower autocorrelation than the observation. With a better modeling of the tails by finding a suitable parameterization, we could obtain more realistic scenarios and expect flatter MTD rank histograms.

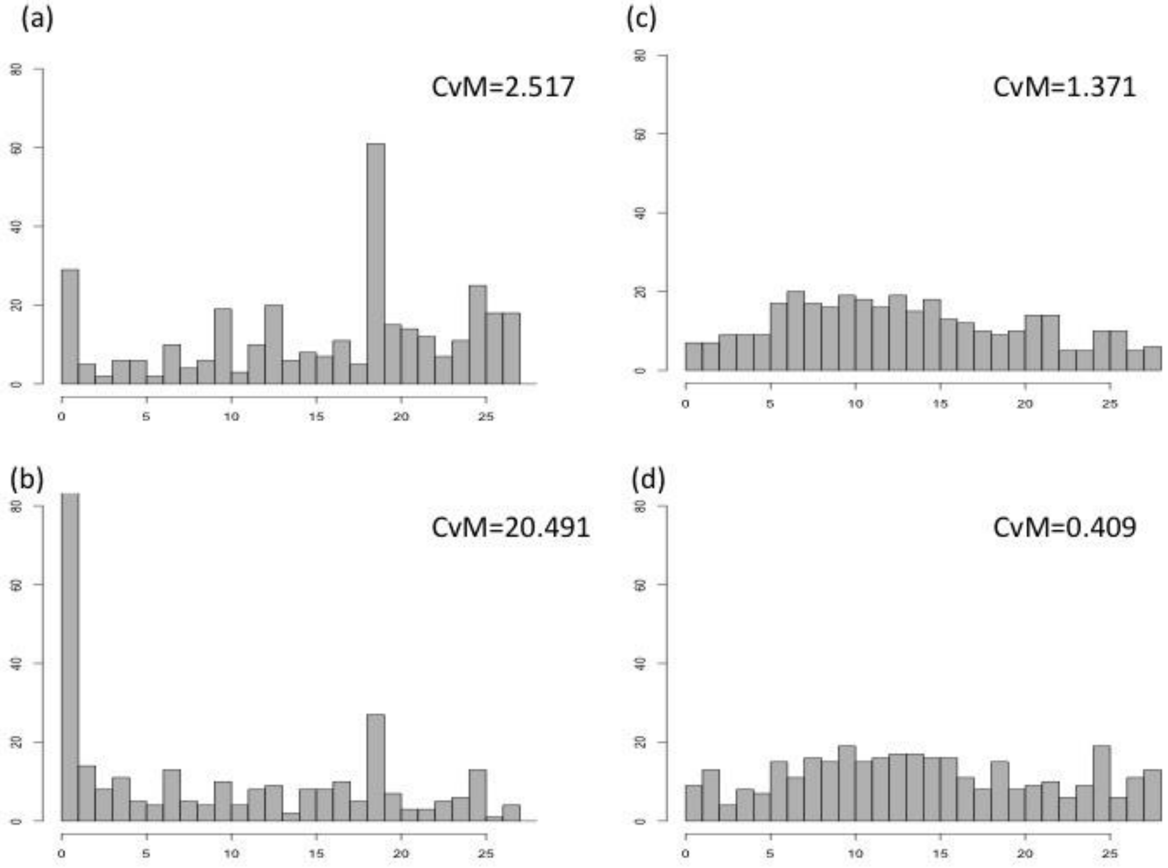


Figure 2.14 Mass transportation distance rank histograms for scenarios generated by (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) QR(1,1) and (d) QR(4,3).

Fig. 2.15 shows how the MTD rank histogram differs from the MST rank histogram for scenarios with unequal probabilities. Results are shown for the EPI(0.1, 0.9) scenarios without de-biasing or scaling.

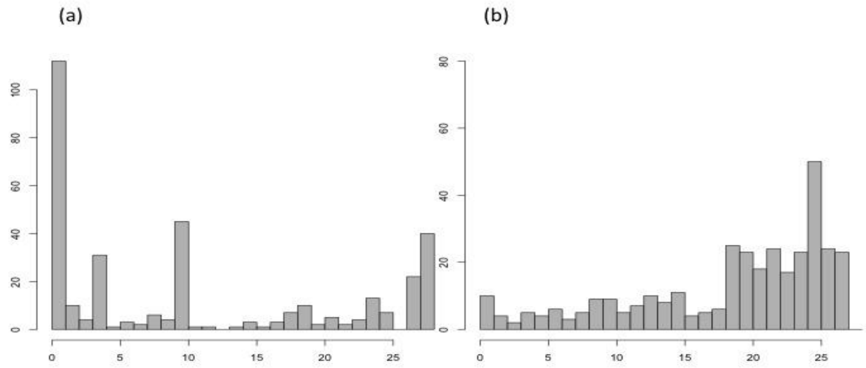


Figure 2.15 (a) MST rank histogram when the scenario probabilities are not considered, (b) MTD rank histogram incorporating scenario probabilities.

Average daily Brier scores for all events and all parameters tested are fairly low for all of the scenario sets, as shown in Table 2.4, and quite similar across the scenario generation methods. For the shortest duration of one hour, the epi-spline scenarios have lower scores but these events are very rare overall. For event 9 (20% ramp down within 6 hours), although the average daily Brier score is slightly higher for scenario set generated by EPI(0.1, 0.9), hourly Brier scores are lower than 0.1 for more than 90% of the hours for the same scenario set, whereas for other scenario sets this proportion is approximately 85%. By changing the parameters, we can capture slight differences among scenario sets. The parameters corresponding to critical events should be determined according to the unit commitment problem results.

Table 2.4 Average daily Brier scores for ramp down and ramp up events with magnitudes ξ = 0.2, 0.4 and durations κ = 1, 3, 6 for scenarios generated by two variants of the epi-spline approximation approach and by three variants of the quantile regression with Gaussian copula approach.

Events:	EPI(0.1, 0.9)	EPI(0.33, 0.66)	QR(1,1)	QR(3,2)	QR(5,3)
1-RampDown(κ=1, ξ=0.2)	0.0015	0.0015	0.0023	0.0025	0.0031
2-RampDown(κ=1, ξ=0.4)	0.0000	0.0000	0.0002	0.0002	0.0004
3-RampUp(κ=1, ξ=0.2)	0.0046	0.0046	0.0050	0.0052	0.0057
4-RampUp(κ=1, ξ=0.4)	0.0001	0.0001	0.0003	0.0004	0.0005
5-RampDown(κ=3, ξ=0.2)	0.0335	0.0325	0.0320	0.0314	0.0321
6-RampDown(κ=3, ξ=0.4)	0.0024	0.0024	0.0029	0.0030	0.0030
7-RampUp(κ=3, ξ=0.2)	0.0452	0.0433	0.0398	0.0401	0.0402
8-RampUp(κ=3, ξ=0.4)	0.0066	0.0064	0.0064	0.0067	0.0069
9-RampDown(κ=6, ξ=0.2)	0.0645	0.0595	0.0614	0.0615	0.0614
10-RampDown(κ=6, ξ=0.4)	0.0133	0.0131	0.0140	0.0143	0.0142
11-RampUp(κ=6, ξ=0.2)	0.0641	0.0601	0.0593	0.0584	0.0602
12-RampUp(κ=6, ξ=0.4)	0.0322	0.0312	0.0297	0.0303	0.0300

The results in this section show the value of employing multiple verification metrics to assess different characteristics of scenarios. The energy score may be appealing as a single number but its emphasis on sharpness could introduce risk. For example, although the lowest ES is obtained from the EPI(0.33, 0.66), the resulting MTD rank histogram is not flat which means it is not as reliable as the other variants of approaches. Thus, we predict that this variant may give the highest cost in SUC problem among all of the scenario sets. The MTD rank histogram identifies whether both variance and autocorrelation in the scenarios match the observations. QR(4,3) is expected to result in slightly lower cost compared to QR(1,1) because its MTD rank histogram is flatter. Brier scores may be very useful but their relative values depend on the definition of critical events. Because high impact events such as steep ramps occur only rarely, the usefulness of statistical assessments may be limited.

2.5 Conclusions

High quality short-term wind power scenarios are very important for achieving cost savings by stochastic unit commitment. Aiming to assess the quality of probabilistic scenarios for wind power trajectories, we employed some existing verification approaches and introduced a mass transportation distance rank histogram to assess calibration of unequally likely scenarios. We applied them to scenario sets that were generated by two very different approaches, one of which produces unequally likely scenarios. On-going research focuses on finding relationships between the verification approach results and unit commitment problem results. We expect MTD rank histograms to predict SUC cost savings better than the ES, because reliability is a more crucial property of wind power scenarios than sharpness. Because scenario sets that are too sharp do not adequately describe the uncertainty, we do not view sharpness is one of the most desired properties of wind power scenarios. It appears that EPI(0.1, 0.9) and QR(4,3) are more competitive than the others presented in this paper.

Another way to predict how scenarios may perform in SUC is by examining Brier scores as well. However, it is very important to decide which events should be considered to distinguish between these scenario sets. The events should be chosen in such a way that they can distinguish whether the scenarios capture steep ramps which (a) may result in costly dispatch decisions in the recourse stage of SUC and (b) are present in the observations. Once these critical events are identified, the decomposition of Brier scores into reliability and resolution components can help to distinguish among sets of scenarios. The ramping event parameters to use in these scores should be determined by careful SUC simulations.

Acknowledgements

This material is based on work supported by the ARPA-E initiative of the U.S. Department of Energy under the Green Electricity Network Integration (GENI) program and by the Iowa Energy Center. We are grateful to Andrea Staid and anonymous reviewers for valuable comments.

References

[1] S. Takriti, J. R. Birge, E. Long. A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems*, Vol. 11, No. 3, August 1996.

[2] A. Tuohy, P. Meibom, E. Denny, M. O'Malley. Unit commitment for systems with significant wind penetration. *IEEE Transactions on Power Systems*, Vol. 24, No. 2, May 2009.

[3] A. Papavasiliou, S. S. Oren, R. P. O'Neill. Reserve requirements for wind power integration: A scenario - based stochastic programming framework. *IEEE Transactions on Power Systems*, Vol, 26, No. 4, November 2011.

[4] K. Cheung, D. Gade, C. Silva-Monroy, S. Ryan, J-P Watson, R. Wets and D. Woodruff. Toward Scalable Stochastic Unit Commitment - Part 2: Assessing Solver Performance. Forthcoming in *Energy Systems*, 2015.

[5] Y. Feng, I. Rios, S. Ryan, K. Spürkel, J-P Watson, R. Wets, and D. Woodruff. Toward Scalable Stochastic Unit Commitment - Part 1: Load Scenario Generation. Forthcoming in *Energy Systems*, 2015.

[6] J. Wang, A. Botterud, R. Bessa, H. Keko, L. Carvalho, D. Issicaba, J. Sumaili, and V. Miranda. Wind power forecasting uncertainty and unit commitment. *Applied Energy*, 88(11):4014-4023, 2011.

[7] P. Pinson, H. Madsen, A. H. Nielsen, G. Papaefthymiou, and B. Klöckl. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12(1):51-62, 2009.

[8] J. O. Royset and R. J.-B. Wets. From data to assessments and decisions: Epi-spline Technology. INFORMS: *TutORials in Operations Research*, 2014.

[9] P. Pinson and R. Girard. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96:12–20, 2012.

[10] T. Gneiting, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2):211-235, 2008.

[11] P. Pinson and J. Tastu. Discrimination ability of the energy score. Technical report, Technical University of Denmark, 2013.

[12] D. S. Wilks. The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, 132, 1329-1340, 2004.

[13] D. Gombos, J. A. Hansen, J. Du, and J. McQueen. Theory and applications of the minimum spanning tree rank histogram. *Monthly Weather Review*, 135, 1490-1505, 2007.

[14] L. A. Smith. Disentangling uncertainty and error: on the predictability of nonlinear systems. *Nonlinear Dynamics and Statistics,* A. E. Mees, Ed., Birkhauer Press, 31-64, 2001.

[15] S. T. Rachev. Probability metrics and the stability of stochastic models. John Wiley, Chichester, UK, 1991.

[16] S.T. Rachev, L. Rüschendorf. *Mass Transportation Problems*, Vol. I and II. Springer-Verlag Berlin, 1998.

[17] J. Dupačová, N. Gröwe-Kuska, and W. Römisch. Scenario reduction in stochastic programming An approach using probability metrics. *Math. Program.,* Ser. A 95: 493–511, 2003.

[18] G. Brier. Verification of forecasts expressed in terms of probability. *Mon Weather Rev,* 78:1-3, 1950.

[19] I. Rios, R J-B Wets, and D.L. Woodruff. Multi-period forecasting and scenario generation with limited data. *Computational Management Science,* 12: 267-295, 2015.

[20] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69, 243–268, 2007.

[21] W. Hsu and A. Murphy. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2: 285-293, 1986.

[22] G. J. Szekely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis,* 93:55-80, 2005.

[23] T. Thorarinsdottir, M. Scheuerer and C. Heinz. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. Forthcoming in *Journal of Computational and Graphical Statistics*, 2015.

[24] J. O. Royset and R. J.-B. Wets. Nonparametric density estimation via exponential epi-eplines: Fusion of soft and hard information. Technical report, Naval Postgraduate School, 2013.

[25] A.T. Al-Awami, and M.A. El-Sharkawi. Statistical characterization of wind power output for a given wind power forecast. *North American Power Symposium (NAPS)*, 2009

[26] H. Bludszuweit, J.A. Dominguez-Navarro and A. Llombart. Statistical Analysis of Wind Power Forecast Error. *IEEE Transactions on Power Systems*, vol. 23, no. 3 983-991, 2008.

[27] S. Bofinger, A. Luig and H. G. Beyer. Qualification of wind power forecasts. *in Proc. Global Wind Power Conf.*, Paris, France, 2002.

[28] http://www.bpa.gov/Projects/Initiatives/Wind/Documents/20140625-BPA-Super-Forecast-Methodology.pdf, viewed 12 March 2015.

[29] http://transmission.bpa.gov/Business/Operations/Wind/default.aspx, viewed 12 March 2015.

[30]http://www.bpa.gov/Projects/Initiatives/Wind/Pages/Wind-Power-Forecasting-Data.aspx, viewed 12 March 2015.

[31]http://transmission.bpa.gov/Business/Operations/Wind/WIND_InstalledCapacity_DATA.pdf viewed 12 March 2015.

[32] T. M. Hamill. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, 14, 155-167, 1999.

[33] N. J. Higham. Computing the nearest correlation matrix – A problem from finance. *MA Journal of Numerical Analysis* 22, 329–343, 2002.

CHAPTER 3: RELIABILITY OF WIND POWER SCENARIOS AND STOCHASTIC UNIT COMMITMENT COST

A paper submitted to the *Energy Systems*

Didem Sari and Sarah M. Ryan

Abstract

Probabilistic wind power scenarios constitute a crucial input for stochastic day-ahead unit commitment in power systems with deep penetration of wind generation. To minimize the expected cost, the scenario time series of wind power amounts available should accurately represent the stochastic process for available wind power as it is estimated on the day ahead. The high computational demands of stochastic programming motivate a search for ways to evaluate scenarios without extensively simulating the stochastic unit commitment procedure. Reliability of wind power scenario sets can be assessed by statistical verification approaches. In this study, we examine the relationship between the statistical evaluation metrics and the results of stochastic unit commitment. Lack of uniformity in a mass transportation distance rank histogram can eliminate scenario sets that might lead to either excessive no-load costs of committed units or high penalty costs for violating energy balance. Event-based metrics can help to predict the cost performance of the remaining scenario sets.

Keywords: Wind power scenarios, Stochastic unit commitment, Reliability, Scenario generation

3.1 Introduction

Wind power generation is taking an increasing share of power generation due to environmental pressures and its low marginal operating cost, which reduces the overall cost of meeting the demand for electrical energy. Thus, its contribution to the total electrical energy production has been increasing rapidly. However, the stochasticity and intermittency of wind requires system operators to schedule the thermal generating units more efficiently in order to accommodate the uncertainty. Inefficient scheduling of generating units may negatively affect both cost and reliability.

The unit commitment (UC) problem, typically solved on the day ahead of the operating day, determines the short-term schedule for the thermal generation units to supply the forecasted power demand. Within the decision making process of scheduling and dispatching electric power generation resources, it is intended to minimize the operational costs that include startup, shutdown, and generation costs while respecting technical and security constraints. For systems with a more conventional generation mix, a UC solution that provides an acceptable response can be obtained by committing a fixed amount of reserve capacity that is available to compensate for load forecast errors. However, as the uncertainty due to renewable penetration increases, these UC solutions can no longer guarantee system security.

The deep penetration of renewable energy resources, such as wind and solar power, leads to increased uncertainty in the net load; i.e., the load less the variable generation. The need to account for this increased uncertainty when optimizing the day-ahead generation schedule has led to great interest in stochastic programming (SP) models for UC [1]. In the SP-based UC formulations, probabilistic scenarios are employed for representing the uncertain net load. To achieve a good solution in a stochastic unit commitment (SUC) program we must formulate a finite number of reasonable scenarios for the time series of variable generation availability over the scheduling horizon.

The quality of the SUC solution is directly linked to the quality of wind power scenarios. In the rapidly growing literature on the modeling and computational aspects of SUC, research on scenario evaluation according to the quality of the resulting SUC solutions is very limited. In this paper, we examine how to judge the quality of a scenario generation method and the scenario sets it produces according to this criterion. We focus on SUC with uncertainty from wind generation.

A distinctive feature of the proposed approach is to employ historical observations of the actual wind power time series and the associated day-ahead forecasts over a time period. Using these data, the ideal way to verify wind power scenarios is as follows: For each day in the historical simulation, generate wind power scenarios on the day ahead using the historical data available up to that time and employ them in the SUC problem. Simulate the implementation of the first stage decisions, which are units' on/off schedules, followed by the economic dispatch decisions according to the actual wind power availability for that day. A good scenario generation method should result in low costs in this historical simulation over a long sequence of days. However, the challenging computational complexity of SUC makes this evaluation method cumbersome. Incorporation of a large number of scenarios in large instances requires the repeated solution of large deterministic equivalent mathematical programs. This challenge motivated a search for ways to evaluate wind power scenario sets (and, by inference, the scenario generation methods that produce them) without having to solve multiple instances of the stochastic program.

We propose to evaluate the reliability of wind power scenarios; i.e., the statistical consistency between the probabilistic scenarios and observations [2,3], using statistical metrics that compare the scenarios against historical data. In this paper, we conduct a historical simulation to investigate relationships between the statistical metric values and the simulated commitment and dispatch costs. A positive relationship would lend credence to assessment of scenario generation methods by the statistical metrics, which can be evaluated quickly. We measure reliability using recently proposed metrics [4] and find that lack of reliability corresponds to high costs in SUC. With the proposed evaluation approach, it is possible to test several scenario generation methods, and choose one that produces scenario sets expected to result in low costs when used in the SUC model.

The paper proceeds as follows: a review of related literature is provided in Section 2. The statistical metrics that we use for wind power scenario evaluation are explained in Section 3. The stochastic unit commitment and dispatch problem is described in Section 4 along with our procedure to simulate its implementation over a historical time period. In Section 5, we explain our case study in detail and provide SUC simulation results over week-long historical time periods from each season of a year, as well as statistical metricscomputed over the whole year, for wind power scenario sets generated by two different scenario generation methods including variants within each method. We examine correspondence between the SUC results and wind power scenario assessment tools for each scenario generation variant. Finally, we conclude in Section 6 with a brief summary and outline of future research directions.

3.2 Literature Review

As global wind power capacity increases, the operational planning in power systems becomes more critical to accommodate variability. Deep penetration of wind power increases uncertainty in net load, which requires more sophistication in the short-term scheduling procedures while it is expected to reduce the overall cost of electrical energy production. The effect of wind power generation on the various components of the operating costs such as the costs of producing power, starting up generating units, CO_2 emissions, etc., is quantified [5] to analyze the impact of wind power generation. Early studies examined the impact of significant wind penetration on short-term scheduling in specific regions [6] and explored the possible improvements to be obtained by accounting for the uncertainty in the optimization [7]. More recent efforts have incorporated sub-hourly dispatch in unit commitment procedures [8] and developed methodology to solve the classical economic dispatch problem in the presence of wind power generation while also accounting for generator reliability uncertainty [9].

Unit commitment is an important short-term planning problem for electrical power generation. It is solved over a specific time horizon to determine when to start up or shut down thermal generating units and how to dispatch the online generators to meet system demand while satisfying generation constraints, such as generation limits, ramping limits, and minimum up/down times, so that the overall operation cost is minimized. The traditional approach to incorporating uncertainty in these scheduling processes is to increase the levels of reserve requirements. Ortega and Kirschen discussed the relationship between UC policies and spinning reserve requirements in terms of cost/benefit [10]. Ela and O'Malley concluded that power system operators must increase reserve margins to account for the larger uncertainty in the net load that results from the rapid growth in renewable generation [11]. Zhou et al. showed how to improve the performance of power system in terms of cost and reliability by scheduling energy and operating reserves that accommodate the wind power forecast uncertainty [12].

By solving SUC problems with probabilistic scenarios for the wind power trajectory, implicit rather than fixed reserve limits are imposed to maintain system reliability. To our knowledge, the first application of stochastic programming to unit commitment was intended to manage uncertainty in demands [13]. The inherent uncertainty and variability of renewable generation revived interest in SUC [14-17], as an alternative to depending on traditional pre-determined reserve limits. Bouffard and Galiana [18] developed a SUC model with a focus on system security. A SUC formulation including reserve requirements was proposed by Ruiz et al. [19] and the results of this combined approach were compared with those of the traditional approach for the efficient management of uncertainty. SUC produces more robust schedules that are better at meeting load and have lower expected costs. Wang et al. included sub-hourly constraints in a SUC model with probabilistic scenarios for wind power [20] while Quan et al. used scenarios to represent not only the uncertainties due to the renewable energy sources, such as wind and solar, but also generator outages [21]. By solving SUC problems where probabilistic scenarios represent the wind power trajectories, cost savings can be achieved in systems with deep penetration of wind power [7,19,22-24].

For an effective stochastic planning approach, the scenario time series of wind power should accurately represent the stochastic process for available wind power. It is critical for a wind power scenario set to follow the observed time series in characteristics such as the levels of wind power available at different time points, the correlations among these levels, and the presence and severity of ramps. Morales et al. proposed a methodology to generate wind speed scenarios for use in stochastic programming decision models [25]. Pinson et al. generated statistical short-term wind power production scenarios [26] and employed statistical metrics to evaluate them [27].

Previous research has identified some statistical metrics that can distinguish between scenario sets. Minimum spanning tree rank histograms are employed for evaluating the reliability of ensemble forecasts [28-30] and equally likely scenarios [27]. The ability of scenarios to represent some critical events that can have an impact on unit commitment and dispatch costs can be assessed by event-based verification. Pinson and Girard defined significant gradient and long lasting events and evaluated sets of equally likely wind scenarios according to those events [27]. Brier scores [31] were used to measure the wind power scenarios' ability to capture the critical events. A mass transportation distance rank histogram informs on the reliability of a scenario set where different probabilities of occurrences can be incorporated [4].

Previous SUC research has focused on improving the mathematical formulation, developing solution approaches to decrease the optimality gap, and devising various scenario reduction techniques to decrease the solution time. Various alternative formulations for unit commitment under uncertainty have been proposed to reduce the computation times [32-34]. Moreover, scenario reduction techniques that are specified to SUC are proposed to decrease the computational demands to a degree [35,36]. Research that considers assessing the scenarios and comparing different scenario sets' performances according to SUC results is very limited. One recent study analyzed different wind power point forecasts by employing them in deterministic unit commitment and comparing with the results of SUC employing wind power scenarios [24]. Our aim is to verify wind power scenario generation methods that are used in SUC problems and compare them according to the costs of unit commitment and dispatch. Although advanced methods are applied to SUC problems to reduce the computational effort [23,37-39], a simulation study to compare a different scenario sets over a historical time period remains computationally very demanding. Therefore, we are interested in ways to evaluate a scenario generation method (or its output) without extensively simulating the stochastic unit commitment procedure. The relationships between wind power scenario assessment metrics and SUC cost components have not been identified so far.

The contribution of this paper is to examine the relationships between wind power scenario assessment metrics [4] and the resulting stochastic unit commitment and dispatch costs. To compare among a number of scenario sets that have been generated by different methods, we search for some statistical metrics that correlate well with the scenario sets' performance in SUC. In this manner, we can distinguish among the scenario sets according to their effect on SUC solutions and, without extensively simulating the SUC, choose a scenario generation method that is expected to yield low costs.

3.3 Statistical metrics for wind power scenario evaluation

Our statistical verification tools for quick evaluation of wind power scenarios are modeled after ensemble forecast verification tools in meteorology. The important properties of an ensemble forecast are reliability, sharpness, skill, and the ability to mimic specific characteristics of the stochastic processes. We focus on how to measure the reliability, which is the degree to which the scenarios and the observations share the same distribution, and the scenarios' ability to capture critical properties of the stochastic process. Ensemble forecast sharpness represents the concentration of the forecasts. From the stochastic optimization perspective, the sharper the set of wind power scenarios, the less uncertainty is represented for consideration in decision making. Thus, we conjecture that sharpness and skill, which encompasses both reliability and sharpness, may not be as appropriate for evaluating scenarios. The forecasts that compose an ensembleare assumed to have equal probabilities of occurrence, whereas scenarios do not have to be equally likely. Thus, the verification tools for ensemble forecasts must be modified to incorporate unequal probabilities. In this study we employ a new reliability metric, the mass transportation distance rank histogram recently proposed[4], as well as event-based verification that is modified to evaluate the combined ability of unequally likely scenarios to represent the pre-defined critical events. These two statistical evaluation metrics are utilized to quantify whether the scenario set possesses desirable properties that are expected to lead to a lower cost in stochastic unit commitment. Define the following notation:

$y_d^0 = \{y_{h,d}^0\}$: observed wind power in hour h=1,…,H on day d =1,…, D

$y_d^s = \{y_{h,d}^s\}$: wind power in hour h=1,…,H on day d =1,…, D, in scenario s = 1,…, S

y_d^{0*}: observed time trajectory on day d, scaled by dividing the wind power levels by the installed capacity.

y_d^{s*}: scaled time trajectory on day d in scenario s.

$y_d^{s\circ}$: de-biased wind power trajectory on day d in scenario s

z_d^0: observation on day *d* standardized according to the Mahalanobis transformation

z_d^s: Mahalanobis-transformed wind power trajectory on day *d* in scenario *s*

p_d^s: probability of occurrence of scenario *s* on day *d*

3.3.1 Mass transportation distance rank histogram

In meteorology and climate science, a minimum spanning tree (MST) rank histogram is used to verify the reliability of multidimensional ensemble forecasts. It is based on the idea that “reliability can be measured by the degree to which the ensemble forecast members and observation are random samples from the same probability density function” [30]. For stochastic programming, we judge a scenario set to be reliable if the probability of event occurrence according to the scenarios matches the relative frequency that event’s occurrence in observations [4]. The MST rank histogram quantifies the reliability of ensemble forecasts where the ensemble members are equally likely but does not accommodate unequal probabilities.

Motivated by the widespread use of the Wasserstein metric in scenario generation and reduction procedures, a mass transportation distance (MTD) rank histogram was developed for assessing the reliability of unequally likely scenarios [4]. The MTD rank histogram is able to distinguish between sets of scenarios that are more or less reliable according to their bias, variability and autocorrelation. The MTD between two discrete distributions is the minimum cost of transporting the probability from one distribution to the other, where cost is proportional to the distance between supporting points of the distributions [40,41]. The MTD rank histogram behaves similarly to the MST rank histogram [29,30] when applied to equally likely scenarios. Its construction is as follows [4]:

(d) Scale the set $\left\{ y_d^0, y_d^1, \ldots, y_d^s \right\}$ to obtain y_d^{0*} and $y_d^{1*}, \ldots, y_d^{s*}$.

(e) Find the MTD for the observation, l_0', which is the distance from the set of scenarios $\left\{ y_d^{k*} : k \in \{1, \ldots, S\} \right\}$ to the observation y_d^{0*}. Then compute the MTD for each scenario, l_j', $j = 1, \ldots, S$, from the set $\left\{ y_d^{k*} : k \in \{0, \ldots, S\} \setminus \{j\} \right\}$ to y_d^j. When computing l_j', assign the probability of scenario y_d^{j*}, which is p_d^j, to the observation y_d^{0*}.

(f) Find the MTD rank, *r*, of the observation MTD l_0', when $l_0', l_1', \ldots, l_s'$ are ordered from largest to smallest. It is an integer between 1 and $S+1$.

Simulation studies demonstrated that MTD rank histograms display a downward trend for an under-dispersed ensemble of scenarios and an upward trend for an over-dispersed ensemble [4]. Flat histograms result when the observation and scenarios are drawn from the same distribution. Bias over-populates the small ranks similarly as under-dispersion. However, high variance in the scenarios can compensate for bias and result in misleadingly flat histograms. To prevent misdiagnosis, the scenarios should be de-biased before constructing MTD rank histograms, according to the following formula:

$$y_{h,d}^{s\circ} = y_{h,d}^{s*} - \frac{1}{D}\sum_{d=1}^{D}\left(\frac{1}{S}\sum_{s=1}^{S} y_{h,d}^{s*} - y_{h,d}^{0*}\right), \text{ for } h = 1, \ldots, H$$

In the context of wind power, we are particularly interested in assessing whether the autocorrelation of scenarios, as a way of describing their temporal smoothness, matches that of observations. Simulation studies were conducted to examine the behavior of MTD rank histogram according to autocorrelation [4]. To de-correlate the data and equalize variances of the marginal distributions, the data were standardized according to the Mahalanobis transformation. The Mahalanobis transformation employs the sample covariance matrix:

$$\mathrm{S}_{scen} = \frac{1}{S}\left[\left(y_d^{0*} - \overline{y}_d^{scen}\right)\left(y_d^{0*} - \overline{y}_d^{scen}\right)^{\mathrm{T}} + \sum_{s=1}^{S}\left(y_d^{s*} - \overline{y}_d^{scen}\right)\left(y_d^{s*} - \overline{y}_d^{scen}\right)^{\mathrm{T}}\right],$$

where

$$\overline{y}_d^{scen} = \frac{1}{S+1}\left(y_d^{0*} + \sum_{s=1}^{S} y_d^{s*}\right)$$

The transformation is a multi-dimensional extension of standardization by subtracting the mean and dividing by the standard deviation:

$$z_d^0 = \mathrm{S}_{scen}^{-1/2}\left(y_d^{0*} - \overline{y}_d^{scen}\right),$$

$$z_d^s = \mathrm{S}_{scen}^{-1/2}\left(y_d^{s*} - \overline{y}_d^{scen}\right)$$

where $\mathrm{S}_{scen}^{-1/2} = D\Lambda^{-1/2}D^{\mathrm{T}}$, D is the matrix whose columns are the eigenvectors of S_{scen}, and $\Lambda^{-1/2}$ is the diagonal matrix containing the reciprocals of the square roots of the corresponding eigenvalues [29].

For over-dispersed scenarios, as the observation autocorrelation decreases, the histogram becomes flatter; however, an upward trend can still be observed. For under-dispersed scenarios, a downward trend is observed for all levels of autocorrelation of the observation but it is less pronounced when the observation autocorrelation is high. If the scenarios and observation have the same autocorrelation and marginal variance, the MTD rank histogram appears to be flat. When the marginal variances of scenarios and observation are the same, the difference between autocorrelations will affect the pattern of the rank histogram. For scenarios with higher (lower) autocorrelation level than the observation, a sloping downward (upward) trend is observed. If we generate scenarios with heterogeneous autocorrelation levels, we observe a hill-shaped MTD rank histogram. This occurs because the presence of both much more and much less smooth scenarios than the observation widens the range of mass transportation distances among scenarios. Thus, the MTD from the scenarios to the observation will fall in the middle rank frequently. Overpopulation of the middle ranks results in a hill-shaped MTD rank histogram that is skewed according to the proportions of scenarios with high and low autocorrelation. Certain combinations of over-dispersion and weak correlation can result in a deceptively flat histogram. This is a limitation of both MTD and MST rank histograms. In summary, the shape of the MTD rank histogram closely corresponds to that of the MST rank histogram when applied to equally likely scenarios. It can also be used to diagnose higher, lower, and mixed levels of autocorrelation in the scenarios compared to the observation. The MTD rank histogram is able to diagnose the same problems as the MST rank histogram. Moreover, it can be used for unequally likely scenarios, whereas MST can be applied only if the scenarios are equally likely.

3.3.2 Event based verification

Event-based verification can be used to explore the scenarios' ability to represent some specific characteristics of stochastic processes. The first step is to determine which stochastic process characteristics are critical to capture. The events can then be defined to detect these critical characteristics. For SUC, we define ramp up (down) events as the maximum increase (decrease) in net load being greater than or equal to a threshold ξ, within a duration of κ hours. By changing the parameters ξ and κ, different specific events can be defined [27]. For wind power scenarios, we are particularly interested in ramp down events because an unexpected loss of a significant amount of wind power could trigger the need for expensive peaking generators to be brought into service. An indicator variable, denoted as $1\{\cdot\}$, takes value 1 if the event occurs or 0 otherwise. Ramp events beginning in hour *h* are defined as follows for a given time series:

$$\text{RampUp}(y_d;h,\kappa,\xi)=1\left\{\exists\ i\in\{0,1,...,\kappa-1\}\quad \text{s.t.}\quad y_{(h+\kappa),d}-y_{(h+i),d}\geq\xi\right\}$$

$$\text{RampDown}(y_d;h,\kappa,\xi)=1\left\{\exists\ i\in\{0,1,...,\kappa-1\}\quad \text{s.t.}\quad y_{(h+i),d}-y_{(h+\kappa),d}\geq\xi\right\}$$

Denoting the parameter set as $\theta=(\kappa,\xi)$, $\text{RampUp}(y_d^0;\theta)$ and $\text{RampDown}(y_d^0;\theta)$ define the ramp up and ramp down events for observed time series on day *d* beginning at time *h* within a time window of length κ. For the scenarios, the event probabilities can be defined mathematically as:

$$\text{P}_{h,d}[\text{RampUp}(y_d^s;\theta)]=\sum_{s=1}^{S}\text{RampUp}(y_d^s;\theta)p_d^s$$

$$\text{P}_{h,d}[\text{RampDown}(y_d^s;\theta)]=\sum_{s=1}^{S}\text{RampDown}(y_d^s;\theta)p_d^s$$

The probability-weighted average of indicator variables for the scenarios takes a value in the interval [0,1]. The Brier score is a strictly proper score to assess these binary situations, which depend on the occurrence and non-occurrence of the event, as applied in [27]. It is computed as the sum of squared distances between the observation indicator and scenario average [31]. An hourly Brier score can be computed as:

$$\mathrm{Bs}(h,d;\mathrm{RampDown}(\theta))_{hourly} = \left(\mathrm{P}_{h,d}[\mathrm{RampDown}(y_d^s;\theta)] - \mathrm{RampDown}(y_d^0;\theta)\right)^2 \qquad \text{for } d = 1,\dots,D$$

whereas we define a daily Brier score as:

$$\mathrm{Bs}(d;\mathrm{RampDown}(\theta))_{daily} = \frac{1}{(H-\kappa)}\sum_{h=1}^{H-k}\mathrm{Bs}(h,d;\mathrm{RampDown}(\theta))_{hourly}.$$

Brier scores measure the degree of correspondence between scenarios and observation based on the event occurrence. They are lower for scenario sets that more accurately reflect the frequency of the event's occurrence.

3.4 Stochastic unit commitment and dispatch problem

Deep penetration of wind power requires more sophistication in operational planning to accommodate variability. One of the most significant short-term planning problems for electric power generation is unit commitment, in which an optimal on-off schedule is found for each thermal generating unit over a given period of time [13]. In the two-stage SUC formulation, the unit commitment decisions are usually made in the first stage and the dispatch decisions are made in the second stage [42]. That is, dispatch decisions are scenario-dependent whereas commitment decisions (except, possibly those of fast-start units) do not depend on the scenarios. The two-stage stochastic program minimizes startup and shutdown costs in the first stage as well as expected generation cost and penalties on load mismatch in the second stage while satisfying operational restrictions over all scenarios. The abstract version of the two-stage model is as follows:

$$f(S) = \min_{v} \; c^{\mathrm{T}} v + \mathrm{Q}(v, S) \qquad (1)$$

$$s.t. \quad Av = b \qquad (2)$$

$$v \text{ binary} \qquad (3)$$

where

$$\mathrm{Q}(v, S) = E_S\left[Q(v, \mathrm{s})\right] \qquad (4)$$

$$Q(v, s) = \min\left\{q_s^T u_s \mid W u_s = h_s - T_s v\right\} \qquad (5)$$

Scenarios, *s*, have a finite discrete distribution. They represent probabilistic time trajectories for wind energy over the scheduling time period. The objective function, represented by equation (1), includes two parts. The first-stage cost related to commitment of units, $c^{\mathrm{T}} v$, includes total startup, shutdown, and no-load costs of committed units. The second-stage cost, $\mathrm{Q}(v, S)$, is the expected value over a given set of scenarios, *S*, which includes the generation cost and penalties on load mismatch given the unit commitments, *v*, from the first stage (4). Equation (2) enforces the minimum up and down time constraints for the binary variable *v*. After realizing each scenario given the commitment of units, formula (5) minimizes generation cost and penalties on load and reserve requirement imbalances. Energy balance, transmission, and ramp rate constraints as well as generation level limitations, etc., related to every concrete scenario are also summarized in the feasible region described by (5). Complete recourse is guaranteed by including slack variables in the energy balance constraints, whose values quantify the load mismatch. A shortage occurs if the sum of energy amounts from dispatching the committed units is less than the net load (load less wind energy) at a specific period, while excess occurs if the sum is greater than the net load.

3.4.1 Simulation procedure over a historical time period

The deterministic equivalent of the stochastic program can be solved in its extensive form as a large-scale mixed-integer program. To assess the performance of wind energy scenarios in an out-of-sample simulation, we solve the stochastic unit commitment problem for a specific day, and then obtain the economic dispatch for the same day with the observed net load using the fixed first stage decision variables v_d^* from the SUC as done in [24]. The simulation procedure over a historical time period is depicted in Figure 3.1:

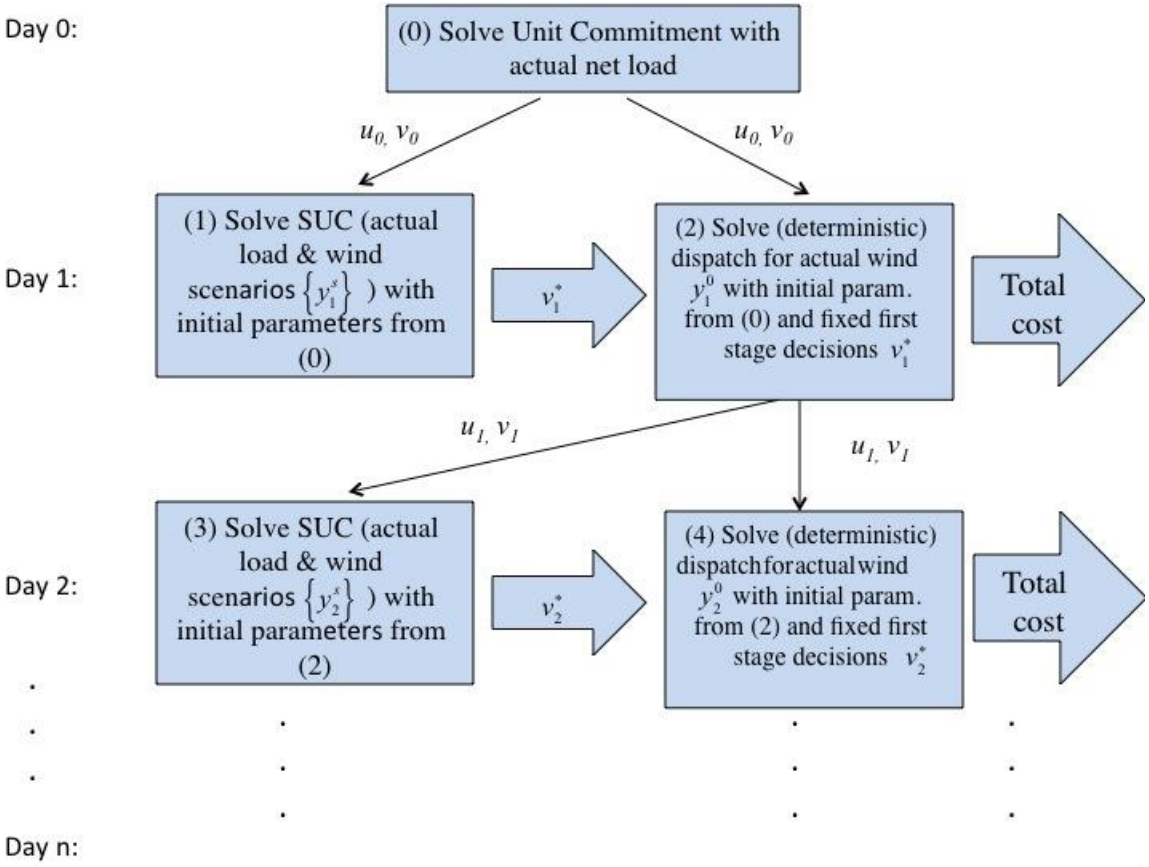

Day 0:
(0) Solve Unit Commitment with actual net load
u_0, v_0
u_0, v_0
Day 1:
(1) Solve SUC (actual load & wind scenarios $\{y_1^s\}$) with initial parameters from (0)
v_1^*
(2) Solve (deterministic) dispatch for actual wind y_1^0 with initial param. from (0) and fixed first stage decisions v_1^*
Total cost
u_1, v_1
u_1, v_1
Day 2:
(3) Solve SUC (actual load & wind scenarios $\{y_2^s\}$) with initial parameters from (2)
v_2^*
(4) Solve (deterministic) dispatchforactualwind y_2^0 with initial param. from (2) and fixed first stage decisions v_2^*
Total cost
Day n:

Figure 3.1 SUC and dispatch simulation procedure over a historical time period

We initialized the unit commitment solution procedure using the commitment at the end of the previous day of the historical time period to set the units' initial on/off states and past durations as well as power generation levels at the beginning of the solution time period. To decide the initial parameters on Day 1 of the planned historical time period, we solved the deterministic unit commitment and dispatch problem with the observed load and observed wind energy for the previous day, which is represented by Day 0 in Figure 3.1. Minimum up and down time constraints for the generation units can affect the next day's initial decisions because if a unit is on (off) at the end of the day, it still must be on (off) the next day until it reaches its minimum up (down) time. Moreover, to satisfy the ramp rate constraints for the first hour of the day, we need previous day's power generation amounts for the last hour. To mitigate end-of-horizon effects, we employed a planning horizon of 36 hours, where we repeated the first 12 hours' net load demand for the last 12 hours. For the first day of the historical period, we solved the stochastic unit commitment problem with probabilistic wind power scenarios y_1^s given the initial parameters from the previous day and obtained the first stage decision variables, v_1^*. We fixed them to their optimal values and solved the economic dispatch problem which is represented by equations (4) and formula (5) with the actual load and observed wind y_1^0 as well as the same initial parameters obtained from the previous day. Because fixed commitments are applied in the economic dispatch problem, the start-up costs and minimum up and down time constraints do not need to be considered. Finally, second stage decision variable u_1 is obtained with actual net load rather than expectation over scenarios. The total costs and results are recorded and the same steps repeated for the remaining days of the historical time period.

This procedure is repeated for using wind power scenarios for each day generated by each of several methods. The hypothesized scenario impacts on cost are as follows: Scenarios that are focused too narrowly (too sharp in the parlance of ensemble forecasts) cause failure to account for the actual risk, and too few low-cost units committed. This may result in starting up additional high cost units or even penalties on load mismatch. If the scenarios are too widely dispersed, the optimization result is too risk averse, and too many units are committed. This may result in excessive no-load cost of committed units. Failure to capture the likelihood of critical events, such as severe down-ramps in wind energy, in the scenarios may likewise result in high dispatch costs.

3.5 Case study

For our case study, we compiled data to represent a recent year in a down-scaled representation of a region. For statistical assessment we used the whole year's worth of scenarios, whereas we arbitrarily choose one week from each season to assess the wind power scenarios' performance according to the SUC simulation results.

3.5.1 The dataset

To generate wind power scenarios we used the day-ahead wind forecast and observation data from the Bonneville Power Administration from 2012/10/01 to 2013/09/31 [43,44]. The days with missing data and/or with wind states considered abnormal are ignored, as documented in detail in [4]. We obtained the load data from Independent System Operator of New England (ISO-NE) [45]. All eight load zones in ISO-NE were treated as a single bus. To focus on the effects of wind power uncertainty we used the observed load and generated probabilistic scenarios only for wind power. Thus, the net load scenarios are obtained by subtracting wind power scenarios from the observed load. A representative subset of 20 generators was selected to keep the computation time manageable. Thus, we scaled the net load scenarios (observed load less wind scenarios) and the observed net load by the ratio of the capacity of the selected generators to the total installed capacity. When simulating the SUC procedure we assume that there is a 20% wind penetration. We imposed severe penalties in the optimization of $1 million and $10 thousand per kWh as penalty costs for shortage and excess, respectively.

3.5.2 Wind power scenario generation

Two different methods were used to generate scenarios. The first one is the quantile regression with Gaussian copula approach [26,4]. For this method, we start by estimating a distribution of day-ahead wind power forecast error based on the historical data after the day-ahead wind power forecast is obtained. For each hour of the day ahead, a quantile regression model estimates the 0.05, 0.10, …, 0.95 quantiles of forecast error based on forecasted wind power generation. Then, by linearly interpolating the quantiles with hypothetical minimum and maximum forecast error, we estimate the distribution of forecast error for each hour. The Gaussian copula method transforms the 24 hourly forecast error distributions into a multivariate Gaussian distribution with covariance of forecast errors of different hours. Thus, we can easily generate forecast error scenarios by projecting Monte Carlo samples from the multivariate Gaussian distribution onto 24 forecast error distributions. Finally the forecast error scenarios are added to the day-ahead forecast to generate day-ahead wind power scenarios (labeled as QR) [4]. Moreover, we have re-modeled the tails by adding another quantile (0.01

and 0.99) to linearly interpolate and truncating remaining tails (<0.01, >0.99). This variant results in slightly smoothed scenarios (labeled as QRnew).

The second scenario generation method is an epi-spline approximation approach [46], and two different variants are generated with cutting probabilities (0, 0.1, 0.9, 1) and (0, 0.33, 0.66 1) [4]. For this method, first error distributions are approximated using exponential epi-splines. Hours within the day are partitioned into intervals by day-part separators (dps; i.e., specific hours of the day). A set of cutting probabilities of error distributions is predefined. Scenarios are generated by the following steps [47]:

1. Cluster forecasts in the training set according to patterns of their left-hand, right-hand and pointwise derivatives at dps hours.
2. Within each cluster, approximate the log error density for each hour with an epi-spline as described in [46]. Numerically integrate to obtain cdfs of the error distributions.
3. Given a forecast of hourly wind power values, identify the cluster to which the forecast belongs at each dps. Compute “skeleton points” for each dps as conditional expected values between the quantiles corresponding to the predefined cutting probabilities.
4. Combine the skeleton points throughout the day.

We use the labels EPI(0.1, 0.9) and EPI(0.33, 0.66) to denote the epi spline scenarios obtained with different cutting points. Each wind power scenario set has 27 scenarios.

3.5.3 Results

The SUC and dispatch problems were solved in their extensive forms by PySP [48] using CPLEX as the MIP solver over the selected historical time periods. Results of the historical time period simulations and statistical metrics are presented in the subsections, respectively. In the plots presented in the following subsections, we use date IDs given in Table 3.1.

Table 3.1 Date IDs and date ranges of the 4 historical time periods used for the SUC simulations

Hist. time period	Date range	Date ID (i=1,…,7)
1	2012/10/17 – 2012/10/23	1_i
2	2013/01/01 – 2013/01/07	2_i
3	2013/04/14 – 2013/04/20	3_i
4	2013/07/07 – 2013/07/13	4_i

3.5.3.1 SUC and dispatch simulation results

Figure 3.2 plots the cumulative costs of our SUC and dispatch simulation for wind power scenarios generated by EPI(0.1, 0.9), EPI(0.33, 0.66), QR and QRnew. As can be seen, EPI(0.1, 0.9) performs the best, whereas EPI(0.33, 0.66) is the worst. And we can observe a slight improvement in QRnew, when we compare it with the QR.

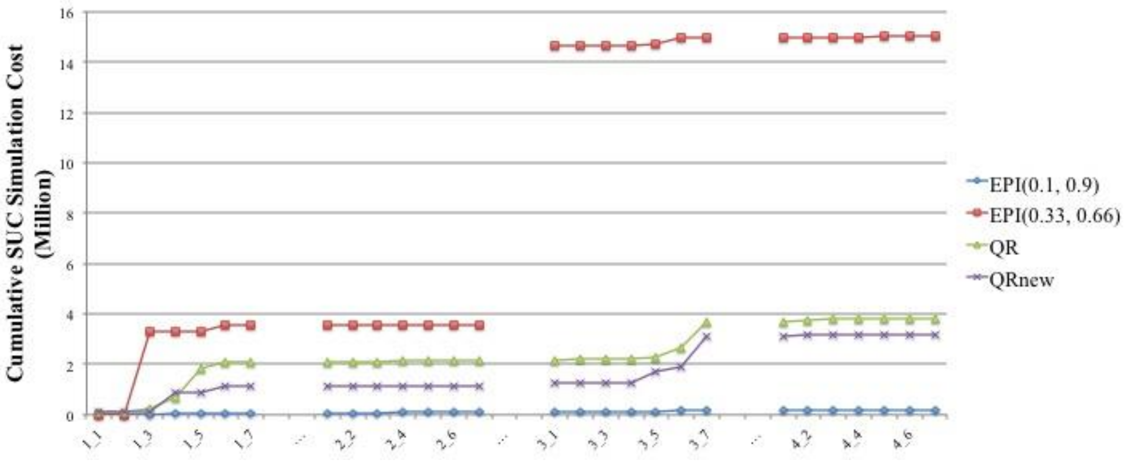


Figure 3.2 Cumulative SUC and dispatch costs over 4 historical time periods

Figures 3.3 and 3.4 show the cumulative deviations from optimal generation levels in megawatt hours incurred by epi-spline and quantile regression scenarios over 4 historical time periods, respectively.

Tables 3.2 and 3.3 shows the dates when there occurs positive and/or negative mismatch by epi-spline and quantile regression scenarios, respectively. The excess and shortage amounts relative to the optimal generation levels are expressed as percentages.

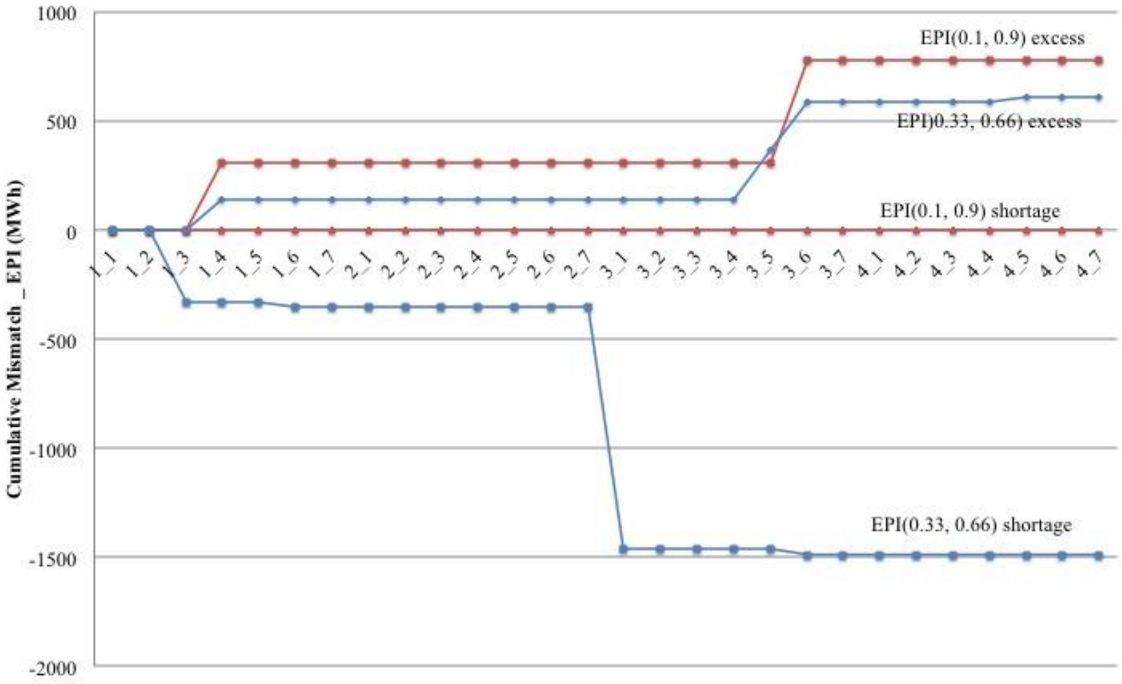


Figure 3.3 Cumulative deviations from optimal generation level occurred by epi-spline scenarios over 4 historical time periods

Table 3.2 Percentages of deviations from the optimal generation levels by Epi spline scenarios.

	EPI (0.1, 0.9)		**EPI (0.33, 0.66)**	
Date	**excess (%)**	**shortage (%)**	**excess (%)**	**shortage (%)**
2012/10/19	0	0	0	0.674
2012/10/20	3.221	0	0.785	0
2012/10/22	0	0	0	0.031
2013/4/14	0	0.005	0	3.121
2013/4/18	0	0	0.779	0
2013/4/19	1.756	0	0.098	0.848
2013/4/20	0.096	0	0	0
2013/7/11	0	0	0.295	0

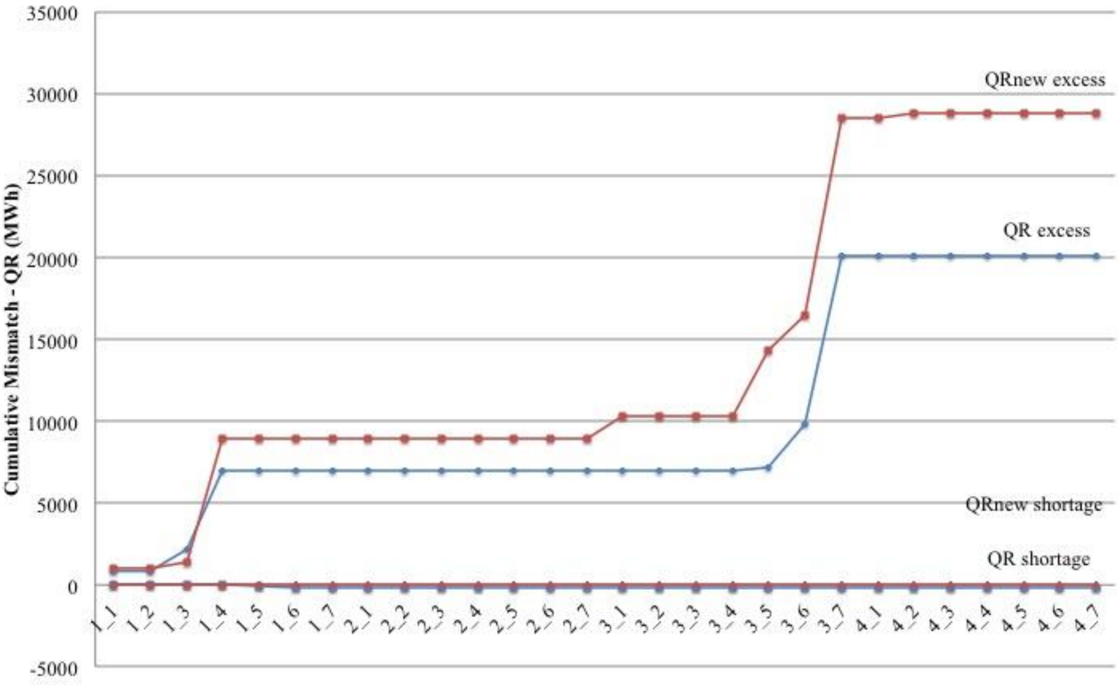


Figure 3.4 Cumulative deviations from optimal generation levels occurred by quantile regression scenarios over 4 planning historical time periods

Table 3.3 Percentages of deviations from the optimal generation levels by quantile regression scenarios.

	QR		**QRnew**	
Date	**excess (%)**	**shortage (%)**	**excess (%)**	**shortage (%)**
2012/10/17	1.489	0	1.790	0
2012/10/19	2.729	0	0.770	0
2012/10/20	26.880	0	42.750	0
2012/10/21	0	1.511	0	0
2012/10/22	0	0.045	0	0.310
2013/4/14	0	0	3.990	0
2013/4/15	0	0.012	0	0
2013/4/18	0.604	0	13.660	0
2013/4/19	9.909	0.044	8.430	0
2013/4/20	245	0	288	0
2013/7/8	0	0.011	0.330	0

3.5.3.2 Statistical metric results

Figure 3.5 shows the MTD rank histograms of wind power scenarios after de-biasing and scaling. We use the Cramér-von Mises goodness of fit test to quantify the uniformity of the resulting MTD rank histograms because it is sensitive to skewed rank histograms.

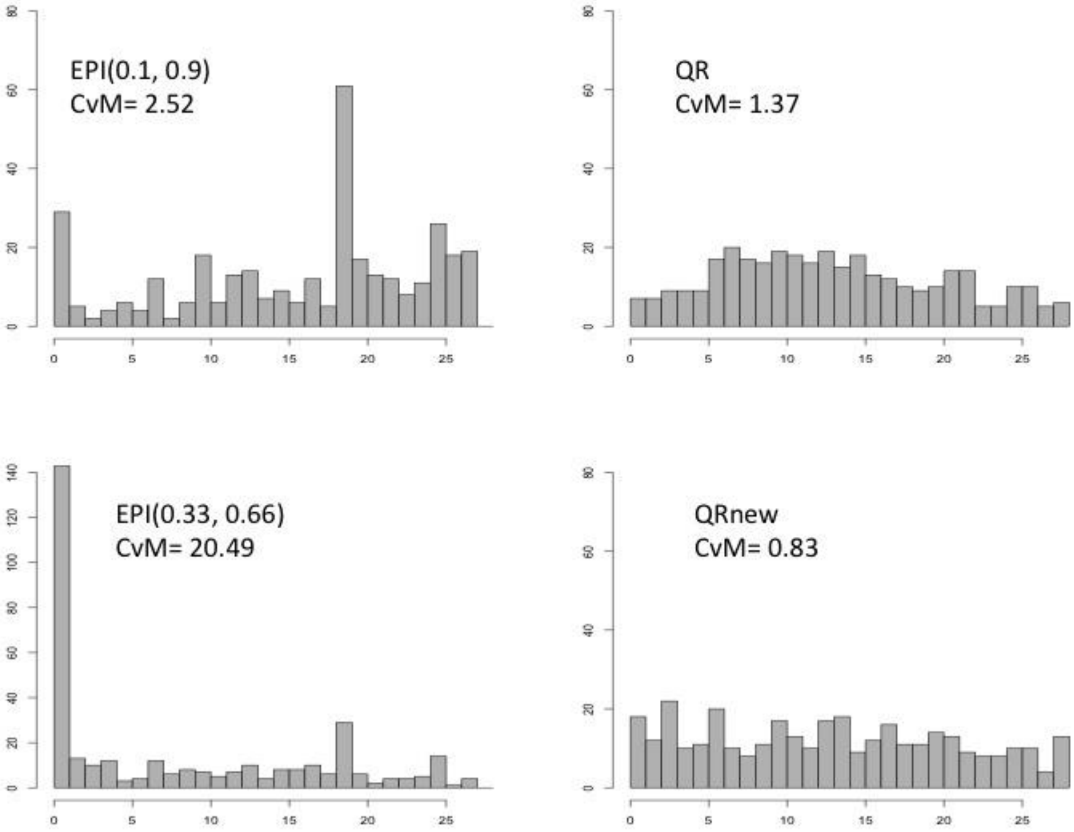


Figure 3.5 Mass transportation distance rank histograms for scenarios EPI(0.1, 0.9), EPI(0.33, 0.66), QR, and QRnew

We assessed the scenarios according to the RampDown event with 2 different sets of parameters. The average daily Brier scores are represented in Table 3.4 and the plots in Figure 3.6 show the average hourly Brier scores and average hourly loads.

Table 3.4 Average daily Brier scores for RampDown event with two different parameters for scenarios EPI(0.1, 0.9), EPI(0.33, 0.66), QR, and QRnew.

Events:	EPI(0.1, 0.9)	EPI(0.33, 0.66)	QR	QRnew
1-RampDown(κ=1, ξ=0.2)	0.0015	0.0015	0.0023	0.0018
2-RampDown(κ=1, ξ=0.4)	0.0000	0.0000	0.0002	3.3e-06

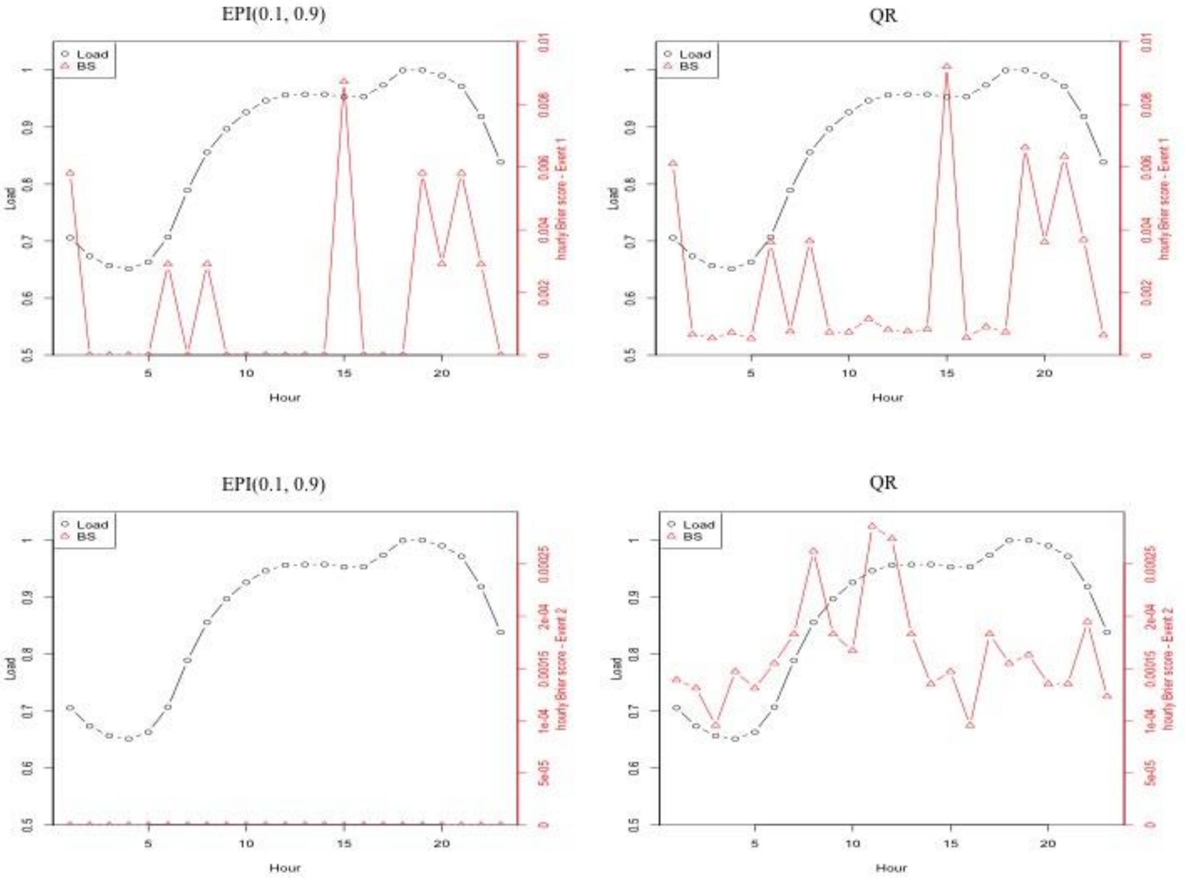


Figure 3.6 Hourly average load and average hourly Brier scores for Event 1 and 2

3.5.4 Discussion

The MTD rank histogram is useful for checking the reliability of scenarios according to their bias, variability, and autocorrelation, as mentioned earlier. Apparently in Figure 3.5 the smallest rank is over-populated in the histogram of EPI(0.33, 0.66), which indicates under-dispersion. This scenario set prevents the optimization from accounting for the actual risk due to the inherent uncertainty in wind. The high difference in SUC and dispatch cost in Figure 3.2 is due to the high penalties assigned to positive mismatch (shortage) in load and startup costs for additional high-cost units. The largest proportion of the cost is due to the unsatisfied demand, which may happen when the observed wind power availability is lower than all of the wind power scenario trajectories at some time period.

The cumulative deviations from optimal generation levels and percentages of deviations incurred by epi-spline scenarios are represented in Figure 3.3 and Table 3.2, respectively. EPI(0.33, 0.66) results in higher and more frequent shortage than the other scenario sets. This is a result of under-dispersion as indicated by the overpopulation of smallest ranks in MTD rank histogram in Figure 3.5. We conjecture that using EPI(0.33, 0.66) scenarios will result in the highest cost over the whole year. As explained and shown by simulation studies in [4], heterogeneous autocorrelation levels in scenarios cause rank histograms to be hill-shaped, as observed in the histogram for QR in Figure 3.5. This is one result of having both very wildly varying and smooth scenarios. Optimization is risk averse with those scenarios and as a result too many low-cost units may be committed, and excessive no-load costs of committed units occur. Moreover, too many committed units will cause penalty costs due to excess because of the minimum power generation limit constraints of the units as seen in Figure 3.4 and Table 3.3. Thus, the penalty costs for excess are higher and occur more frequently for the quantile regression scenarios than the epi-spline scenarios (Tables 3.2 and 3.3). Moreover, even if we ignore all penalty costs due to the mismatch in load, we still observe that quantile regression scenarios result in higher costs than do the epi-spline scenarios. After better modeling the tails in the quantile regression scenario generation method, we obtained slightly smoothed scenarios. This is indicated by a flatter histogram as seen in Figure 3.5 (QRnew). Eliminating very wild scenarios results in slightly lower costs in SUC in Figure 3.2. However, we still observe some penalty costs due to shortage in demand in all of the variants of the QR scenarios (Figure 3.3 and Table 3.3). The shortage is not because of the under-dispersion as in EPI(0.33, 0.66), but because of the sampling behavior.

As explained and shown with the simulation studies in [4], under-dispersion, over-dispersion, and differences in autocorrelation levels affect the skewness of the rank histogram, whereas heterogeneous autocorrelation levels in scenario set overpopulate the middle of the rank histogram. Moreover, some combinations of all these statistics may result in a misleadingly flat histogram, which is a limitation of MTD rank histogram. It would not be valid to assert that the wind power scenario set with flattest MTD rank histogram would perform the best in a SUC and dispatch solution procedure over a historical time period. However, we can eliminate the scenario sets having right-skewed and hill-shaped rank histograms because the majority of the costs occur because of under-dispersion (which result in penalties due to positive mismatch in load) and heterogeneous autocorrelation levels in the scenario set (which result in committing too many units, excessive no-load costs and penalties due to negative mismatch in load). In our case study, we can choose EPI(0.1, 0.9) over

EPI(0.33, 0.66) and QRnew over QR. The MTD rank histogram seems to be better able to distinguish among the variants of each scenario generation methods than across the methods.

Table 3.4 shows the average daily Brier scores for RampDown events with two different parameters. One limitation of Brier score to evaluate wind power scenarios is that it gives very low scores when the scenario sets are too sharp. Since the under-dispersed EPI(0.33, 0.66) scenarios are too sharp, they result in low scores whereas their costs are too high in SUC. However, if the scenario set is not under-dispersed, the Brier scores of RampDown events are very successful to catch the differences over the scenarios. As can be seen from Table 3.4, QR has the highest Brier score whereas EPI(0.1, 0.9) has the lowest. In Figure 3.6 we plot hourly average load and average hourly Brier scores according to the events shown in Table 3.4 for the wind power scenario sets that have the highest and lowest average daily Brier scores. The load ramps up after hour 9, and the peak load occurs between hours 12 and 21. Thus, the differences among Brier scores of wind power scenarios during those hours are more critical. If the wind power scenarios do not successfully reflect the likelihood of the RampDown event in that time range, expensive peaking generators would be required to satisfy the unexpectedly high net load.

To summarize, we would expect a successful wind power scenario set to first be reliable, which means a good level of correspondence between scenario distribution and observation distribution according to their autocorrelation and variability and, second, to represent the critical events such as RampDown and RampUp with some specific parameters for our SUC and dispatch problem. We recommend to first use the MTD rank histograms to eliminate the wind scenarios that are right-skewed (under-dispersed) and/or hill-shaped (weak correspondence in autocorrelation level). Then compare the remaining scenario sets according the Brier scores of the RampDown event.

3.5.5 Daily comparisons

In this section, for some specific days we plot wind power scenarios generated by two variants of each methods and represent daily SUC and dispatch costs by comparing some cost components to give additional insight.

In Figures 3.7 and 3.8, we plot the wind power scenarios generated by epi-spline approximation on the left and quantile regression with Gaussian copula approach on the right. Wind energy is scaled according to the capacity.

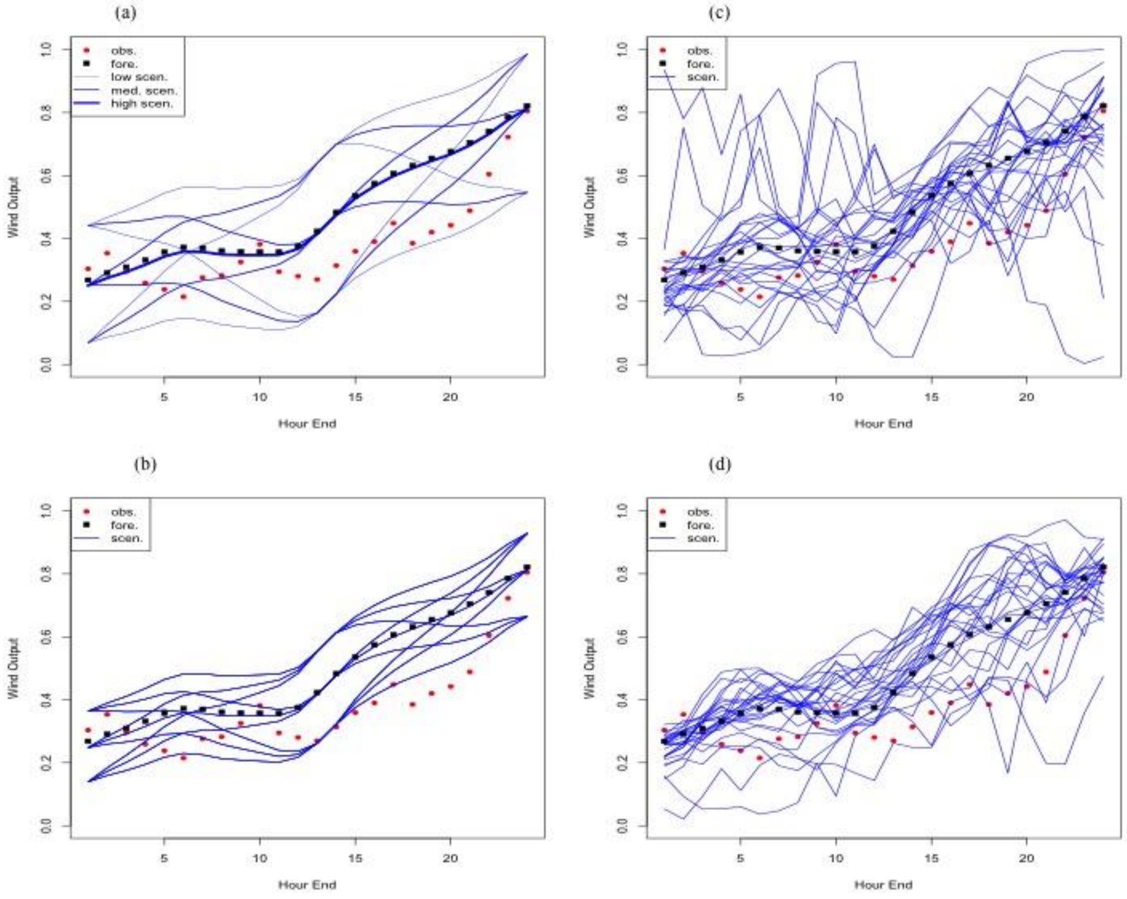


Figure 3.7 Wind power scenarios generated for day 2012/10/19. (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) QR, (d) QRnew

For day 2012/10/19 the SUC costs resulting from using the different wind power scenarios are ordered as EPI(0.33,0.66) > QR > QRnew > EPI(0.1,0.9). The majority of the costs occur because of the penalties for all the scenario sets except EPI (0.1, 0.9). However, EPI(0.33, 0.66) has the highest penalties. No-load costs for EPI scenarios are lower than QR scenarios.

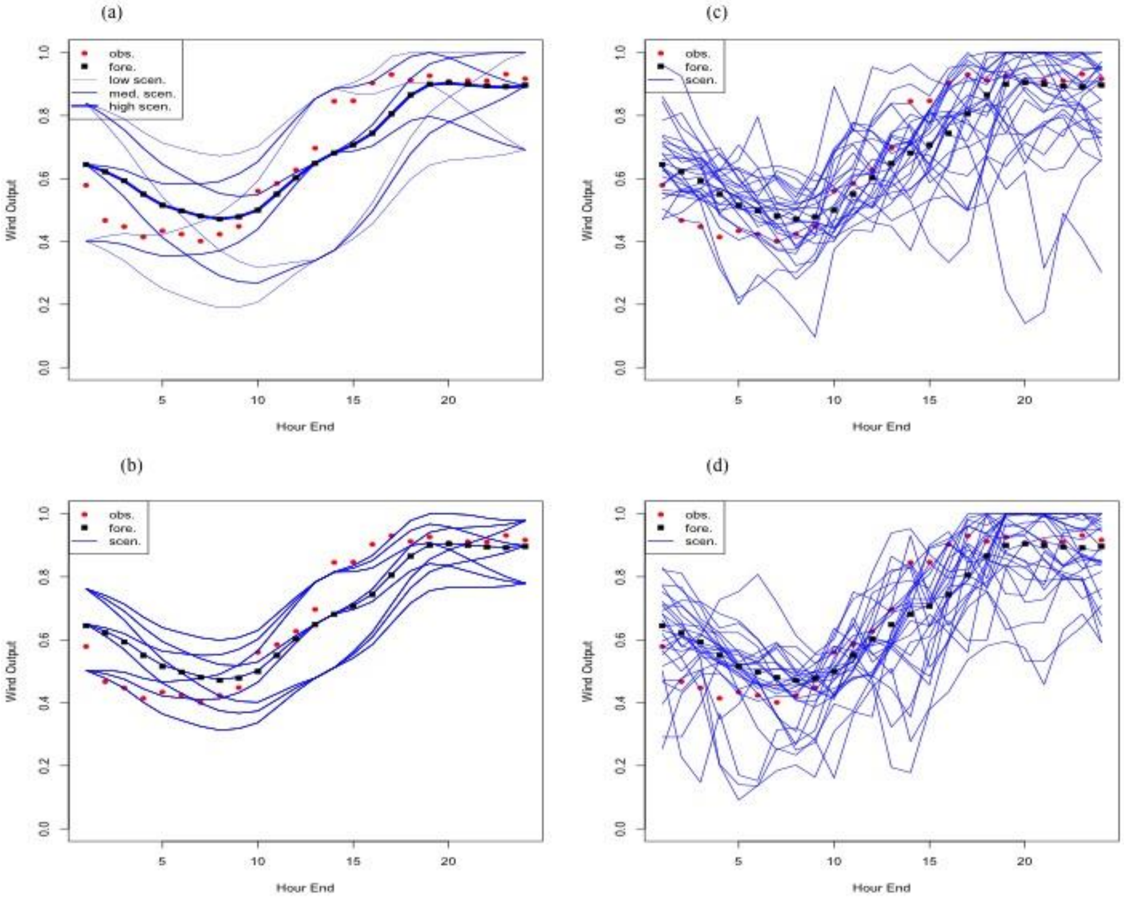


Figure 3.8 Wind power scenarios generated for day 2013/04/19. (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) QR, (d) QRnew

For day 2013/04/19, all of the scenario sets have penalty due to excess, and the amount of excess is ordered as QR > QRnew > EPI(0.1, 0.9) > EPI(0.33, 0.66). Only EPI(0.33, 0.66) and QR caused shortage penalties and the shortage amount is higher for EPI(0.33, 0.66).

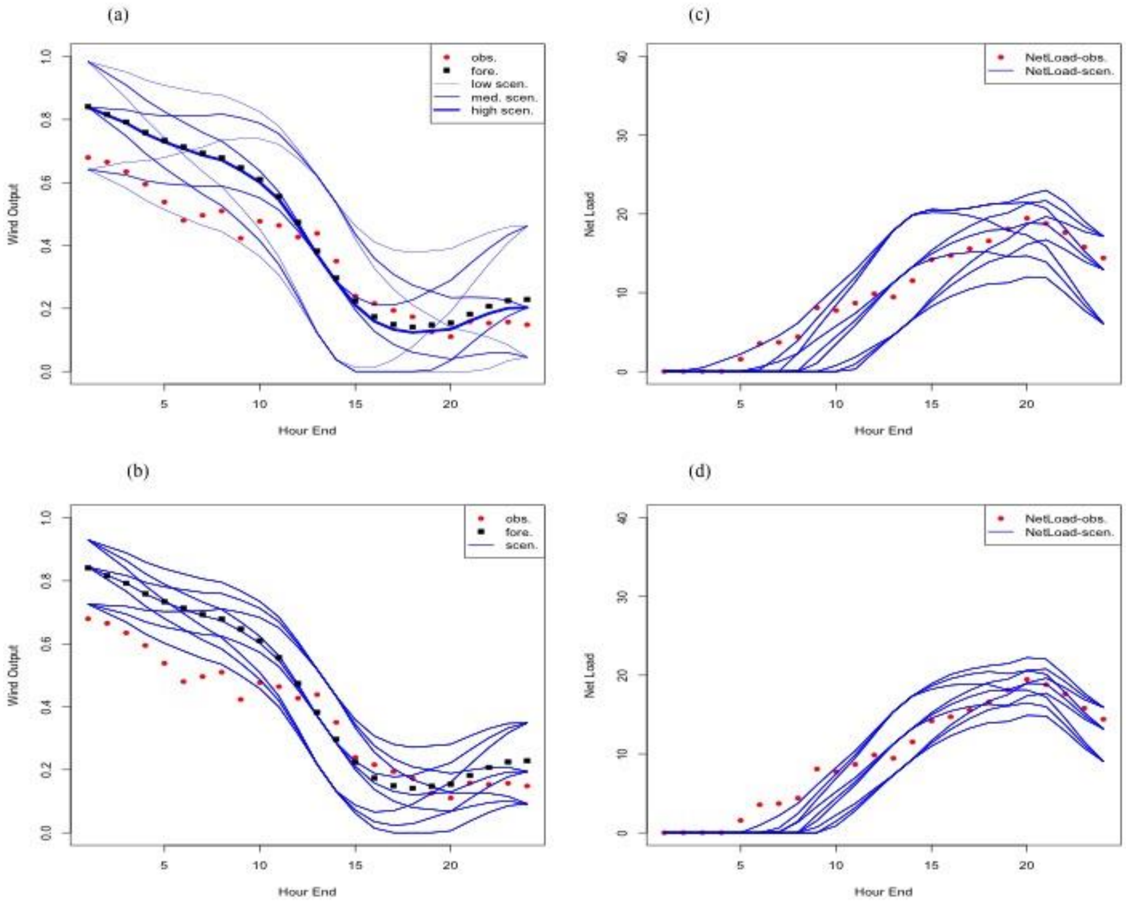


Figure 3.9 Wind power scenarios and net load scenarios for day 2013/04/14. (a) EPI(0.1, 0.9), (b) EPI(0.33, 0.66), (c) Net load scenarios for EPI(0.1, 0.9), (d) Net load scenarios for EPI(.33, 0.66)

Figure 3.9 shows the plots of wind scenarios on the left-hand side and net load (observed load – wind scenario) scenarios after scaling according to the generator capacity and adjusting according to the 20% wind penetration on the right. On day 2013/04/14, no negative mismatch occurred for both of the epi spline wind scenarios. However, there was positive mismatch for both, which corresponded to 0.005% and 3.121% deviation from the optimal generation level for EPI(0.1, 0.9) and EPI(0.33, 0.66), respectively.

3.6 Conclusions

A stochastic unit commitment formulation can achieve cost savings where uncertainty occurs in wind power but the computation time increases with the dimension of the deterministic optimization and the number of scenarios. To facilitate comparison of a considerable number of scenario sets' performance over a long historical time period and choose the best scenario generation method, we employ two statistical metrics: mass transportation distance rank histogram and event based verification. We aim to predict each scenario set's unit commitment performance with these statistical tools. Two different scenario generation methods and their variants have been compared according to their performance in a simulation of the SUC procedure. Thus, we have explored the relationship between these wind power scenario evaluation methods and SUC costs. Using the mass transportation distance rank histogram we can eliminate the scenario sets that might lead to either high no-load costs or high penalty costs due to shortage or excess. Secondly, after defining critical event(s) for the problem we compare the remaining scenarios according to the Brier scores. Both metrics have limitations. For some specific combinations of over-dispersion and weak correlation, the MTD rank histogram appears deceptively flat. Moreover, Brier scores may be very low for under-dispersed and/or sharp scenario sets. According to the results represented in the case study, we can conclude that, of the wind power scenario generation methods and variants tested, scenario set generated by epi-spline approximation approach with cutting probabilities (0, 0.1, 0.9, 1) performs the best in the SUC problem, as could be predicted by its flat MTD rank histogram and low Brier scores for ramp-down events.

REFERENCES

1. Zheng, Q.P.P., Wang, J.H., Liu, A.L.: Stochastic Optimization for Unit Commitment-A Review. IEEE T Power Syst 30(4), 1913-1924 (2015).

2. Gneiting, T., Balabdaoui, F., Raftery, A.E.: Probabilistic forecasts, calibration and sharpness. J Roy Stat Soc B 69, 243-268 (2007).

3. Hsu, W.R., Murphy, A.H.: The Attributes Diagram - a Geometrical Framework for Assessing the Quality of Probability Forecasts. Int J Forecasting 2(3), 285-293 (1986). doi:Doi 10.1016/0169-2070(86)90048-8

4. Sari, D., Lee, Y., Ryan, S., Woodruff, D.: Statistical metrics for assessing the quality of wind power scenarios for stochastic unit commitment. Wind Energy 19(5), 873-893 (2016).

5. Ortega-Vazquez, M.A., Kirschen, D.S.: Assessing the Impact of Wind Power Generation on Operating Costs. IEEE T Smart Grid 1(3), 295-301 (2010).

6. Ummels, B.C., Gibescu, M., Pelgrum, E., Kling, W.L., Brand, A.J.: Impacts of wind power on thermal generation unit commitment and dispatch. IEEE T Energy Conver 22(1), 44-51 (2007).

7. Tuohy, A., Meibom, P., Denny, E., O'Malley, M.: Unit Commitment for Systems With Significant Wind Penetration. IEEE T Power Syst 24(2), 592-601 (2009).

8. Yang, Y.C., Wang, J.H., Guan, X.H., Zhai, Q.Z.: Subhourly unit commitment with feasible energy delivery constraints. Appl Energ 96, 245-252 (2012).

9. Osorio, G.J., Lujano-Rojas, J.M., Matias, J.C.O., Catalao, J.P.S.: A probabilistic approach to solve the economic dispatch problem with intermittent renewable energy sources. Energy 82, 949-959 (2015).

10. Ortega-Vazquez, M.A., Kirschen, D.S.: Optimizing the spinning reserve requirements using a cost/benefit analysis. IEEE T Power Syst 22(1), 24-33 (2007).

11. Ela, E., O'Malley, M.: Studying the Variability and Uncertainty Impacts of Variable Generation at Multiple Timescales. IEEE T Power Syst 27(3), 1324-1333 (2012).

12. Zhou, Z., Botterud, A., Wang, J., Bessa, R.J., Keko, H., Sumaili, J., Miranda, V.: Application of probabilistic wind power forecasting in electricity markets. Wind Energy 16(3), 321-338 (2013).

13. Takriti, S., Birge, J.R., Long, E.: A stochastic model for the unit commitment problem. IEEE T Power Syst 11(3), 1497-1506 (1996).

14. Bakirtzis, E.A., Biskas, P.N., Labridis, D.P., Bakirtzis, A.G.: Multiple Time Resolution Unit Commitment for Short-Term Operations Scheduling Under High Renewable Penetration. IEEE T Power Syst 29(1), 149-159 (2014).

15. Papavasiliou, A., Oren, S.S.: Multiarea Stochastic Unit Commitment for High Wind Penetration in a Transmission Constrained Network. Oper Res 61(3), 578-592 (2013).

16. Wu, H.Y., Shahidehpour, M.: Stochastic SCUC Solution With Variable Wind Energy Using Constrained Ordinal Optimization. IEEE T Sustain Energ 5(2), 379-388 (2014).

17. Madaeni, S.H., Sioshansi, R.: The impacts of stochastic programming and demand response on wind integration. Energy Systems 4(2), 109-124 (2013). doi:10.1007/s12667-012-0068-7

18. Bouffard, F., Galiana, F.D.: Stochastic security for operations planning with significant wind power generation. IEEE T Power Syst 23(2), 306-316 (2008).

19. Ruiz, P.A., Philbrick, C.R., Zak, E., Cheung, K.W., Sauer, P.W.: Uncertainty Management in the Unit Commitment Problem. IEEE T Power Syst 24(2), 642-651 (2009).

20. Wang, J.D., Wang, J.H., Liu, C., Ruiz, J.P.: Stochastic unit commitment with sub-hourly dispatch constraints. Appl Energ 105, 418-422 (2013).

21. Quan, H., Srinivasan, D., Khambadkone, A.M., Khosravi, A.: A computational framework for uncertainty integration in stochastic unit commitment with intermittent renewable energy sources. Appl Energ 152, 71-82 (2015).

22. Ela, E., Milligan, M., O'Malley, M.: A Flexible Power System Operations Simulation Model for Assessing Wind Integration. IEEE Pow Ener Soc Ge (2011).

23. Papavasiliou, A., Oren, S.S., O'Neill, R.P.: Reserve Requirements for Wind Power Integration: A Scenario-Based Stochastic Programming Framework. IEEE T Power Syst 26(4), 2197-2206 (2011).

24. Wang, J., Botterud, A., Bessa, R., Keko, H., Carvalho, L., Issicaba, D., Sumaili, J., Miranda, V.: Wind power forecasting uncertainty and unit commitment. Appl Energ 88(11), 4014-4023 (2011).

25. Morales, J.M., Minguez, R., Conejo, A.J.: A methodology to generate statistically dependent wind speed scenarios. Appl Energ 87(3), 843-855 (2010).

26. Pinson, P., Madsen, H., Nielsen, H.A., Papaefthymiou, G., Klockl, B.: From Probabilistic Forecasts to Statistical Scenarios of Short-term Wind Power Production. Wind Energy 12(1), 51-62 (2009).

27. Pinson, P., Girard, R.: Evaluating the quality of scenarios of short-term wind power generation. Appl Energ 96, 12-20 (2012).

28. Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L., Johnson, N.A.: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17(2), 211-235 (2008).

29. Wilks, D.S.: The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. Mon Weather Rev 132(6), 1329-1340 (2004).

30. Gombos, D., Hansen, J.A., Du, J., McQueen, J.: Theory and applications of the minimum spanning tree rank histogram. Mon Weather Rev 135(4), 1490-1505 (2007).

31. Brier, G.W.: Verification of Forecasts Expressed in Terms of Probability. Mon Weather Rev 78(1), 1-3 (1950). doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

32. Bruninx, K., Dvorkin, Y., Delarue, E., Pandzic, H., D'haeseleer, W., Kirschen, D.S.: Coupling Pumped Hydro Energy Storage With Unit Commitment. IEEE T Sustain Energ 7(2), 786-796 (2016).

33. Siface, D., Vespucci, M.T., Gelmini, A.: Solution of the mixed integer large scale unit commitment problem by means of a continuous Stochastic linear programming model. Energy Systems 5(2), 269-284 (2014). doi:10.1007/s12667-013-0107-z

34. Bruninx, K., Van den Bergh, K., Delarue, E., D'haeseleer, W.: Optimization and Allocation of Spinning Reserves in a Low-Carbon Framework. IEEE T Power Syst 31(2), 872-882 (2016). doi:10.1109/TPWRS.2015.2430282

35. Shukla, A., Singh, S.N.: Clustering based unit commitment with wind power uncertainty. Energ Convers Manage 111, 89-102 (2016).

36. Feng, Y., Ryan, S.M.: Solution sensitivity-based scenario reduction for stochastic unit commitment. Computational Management Science 13(1), 29-62 (2016). doi:10.1007/s10287-014-0220-z

37. Ji, B., Yuan, X.H., Chen, Z.H., Tian, H.: Improved gravitational search algorithm for unit commitment considering uncertainty of wind power. Energy 67, 52-62 (2014).

38. Nasri, A., Kazempour, S.J., Conejo, A.J., Ghandhari, M.: Network-Constrained AC Unit Commitment Under Uncertainty: A Benders' Decomposition Approach. IEEE T Power Syst 31(1), 412-422 (2016).

39. Cheung, K., Gade, D., Silva-Monroy, C., Ryan, S.M., Watson, J.P., Wets, R.J.B., Woodruff, D.L.: Toward scalable stochastic unit commitment Part 2: solver configuration and performance assessment. Energy Syst 6(3), 417-438 (2015). doi:10.1007/s12667-015-0148-6

40. Rachev, S.T.: Probability Metrics and the Stability of Stochastic Models. Wiley, New York (1991)

41. Rachev, S.T., Rüschendorf, L.: Mass Transportation Problems. Probability and Its Applications, vol. 1. Springer-Verlag New York (1998)

42. Feng, Y.H., Rios, I., Ryan, S.M., Spurkel, K., Watson, J.P., Wets, R.J.B., Woodruff, D.L.: Toward scalable stochastic unit commitment. Part 1: load scenario generation. Energy Syst 6(3), 309-329 (2015). doi:10.1007/s12667-015-0146-8

43. http://transmission.bpa.gov/Business/Operations/Wind/default.aspx.

44.http://www.bpa.gov/Projects/Initiatives/Wind/Pages/Wind-Power-Forecasting-Data.aspx. 2016

45. http://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info.

46. Royset JO, W.R.-B.: Nonparametric density estimation via exponential epi-eplines: Fusion of soft and hard information. https://www.math.ucdavis.edu/~rjbw/mypage/Statistics_files/RstW13_xspl.pdf (2013).

47. Rios, I., Wets, R.J.-B., Woodruff, D.L.: Multi-period forecasting and scenario generation with limited data. Computational Management Science 12(2), 267-295 (2015). doi:10.1007/s10287-015-0230-5

48. Watson, J.-P., Woodruff, D.L., Hart, W.E.: PySP: Modeling and solving stochastic programs in Python. https://cfwebprod.sandia.gov/cfdocs/CompResearch/docs/pysp_jnl.pdf (2010)

CHAPTER 4: SCENARIO GENERATION QUALITY ASSESSMENT FOR TWO-STAGE STOCHASTIC PROGRAMS

A paper in preparation for submission

Abstract

In minimization problems with uncertain parameters, cost savings can be achieved by solving stochastic programming (SP) formulations instead of using expected parameter values in a deterministic formulation. To obtain such savings, it is crucial to employ scenarios of high quality. A convincing way to assess the quality of scenarios is to simulate employing the resulting scenarios when solving the SP problem while measuring the costs incurred when the solution is implemented. Simulation studies to assess the scenario quality in this way are computationally very demanding. In this study, we utilize two novel approaches: expected value based and perfect information based scenario generation assessment. With the proposed approaches we can assess the quality of scenario sets without having to repeatedly solve the related SP problem. Instead of comparing scenarios to observations directly, the impact of each scenario in the SP problem is taken into consideration in these approaches. The scenario generation methods that are expected to lead low costs in SP problem can be verified quickly without or before solving the SP problem.

Keywords: Stochastic programming, Two-stage stochastic programs, Scenario generation assessment, Scenario quality

4.1 Introduction

In recent years, stochastic programming (SP) has gained an increasing popularity to solve real world optimization problems which inevitably include uncertainty. It has been applied in various areas, such as energy, finance, transportation, scheduling, production control, and capacity planning. In SP problems, we add stochasticity to models by describing the unknown parameters by their probability distributions. We employ probabilistic scenarios to represent stochasticity. Scenarios are the plausible realizations of the stochastic processes throughout the decision making horizon, with probabilities of their occurrence provided.

The set of scenarios included has a big impact on the solution of a SP problem. Different scenario generation methods model the uncertain events in different ways. The quality of the solution of the SP is directly linked to the quality of scenarios. The number of scenarios needed to accurately represent the uncertainty involved and prevent loss of information is generally large. However, incorporation of a large number of scenarios in the extensive form of the deterministic equivalent leads to huge mathematical programs, which causes high computation times especially if discrete decision variables are present. The decision of whether a scenario set or scenario generation method is of high quality and the process of comparing scenario generation methods over a time horizon is challenging due to the high computation times. Solving the associated SP problem repeatedly for each scenario set, which we would like to compare according to their performance in stochastic problems, can become intractable. Thus, we are searching for approaches to assess scenarios without having to repeatedly and explicitly solve the stochastic problem.

This paper proposes a novel approach for assessing the quality of a scenario generation method or equivalently, the sets of scenarios it produces, using historical outcomes. The quality of scenario sets is related to the quality of solutions obtained when using those scenarios in the SP, as described in Section 4.3. On the other hand, reliability is the statistical consistency between the probabilistic scenarios and observations. Reliability is a prerequisite for quality. We accept scenario sets (and, consequently, a scenario generation method) to be reliable if the observed relative frequency of events represented by a scenario is close to the scenario probability [1]. We assume that observational data on the values of the uncertain parameters are available for a historical time period and that the study period will be similar. Existence of the historical data implies that we are solving a problem that repeats rather than a one-shot design type of problem. The goal is to develop a general approach for quality assessment that could be applied to any two-stage SP problem that repeats.

In our approach, the distances among scenarios and the observed value are measured by fixing first-stage decisions to a common value and computing second-stage objective values. A rank histogram constructed from these distances, motivated by the mass transportation metric used in scenario reduction, can diagnose bias or other defects in the scenario set. The mass transportation distance (MTD) between distributions is the minimum cost of transporting the probability from one distribution to the other, where cost is proportional to the distance between supporting points of the distributions [2, 3]. The method is demonstrated by simulations of stochastic server location and unit commitment instances. Simulation studies show that this approach can distinguish high quality from low quality scenario sets. A case study is provided for the stochastic unit commitment (SUC) problem where a scenario generation method that is expected to lead to low costs when implemented in the SUC and the subsequent dispatch is identified.

The contribution of this paper is to provide an efficacious technique that allows assessment of scenario generation results in the context of a target optimization problem without or before explicitly solving the SP repeatedly. Instead, only deterministic problems corresponding to single scenarios are solved. This method is intended for use when the original problem is difficult to solve, when the problem is repetitive rather than a one-shot design type of problem, and where the historical observations of the uncertain parameters are available for a similar time period.

The paper proceeds as follows: Section 4.2 provides a literature review on scenario reduction and scenario evaluation for SP problems. In Section 4.3. we explain the scenario quality assessment by historical simulation. Proposed generic approach for assessment of scenario generation methods is explained in Section 4.4. In Section 4.5, we describe our motivating models. We present the proposed approaches that are expected value based and perfect information based scenario assessment, along with the simulation studies on motivating models, in Section 4.6 and 4.7, respectively. In Section 4.8, we explain our case study on SUC in detail. We provide the results of scenario assessment approaches using wind power scenarios that are generated by two different scenario generation methods, including variants within each method. Finally, we conclude in Section 4.9 with a brief summary and discussion of research directions.

4.2 Literature Review

Evaluation of scenario generation methods for SP was studied by Kaut and Wallace [4]. They discussed and formulated important properties that a scenario generation method should possess in order to be usable for a given decision model. The optimality gap is defined as the difference between the objective function values at the optimal solutions of the true and the approximated problems. They observed that it is impossible to test the optimality gap in most practical problems because it requires solving the optimization problem with the true values of parameters, which are unknown. In our proposed methodology we assume that the historical outcomes including the observational; i.e., true, data are available, which may be a valid assumption for problems where past statistics are available, such as weather forecasts along with observations for the SUC problem or exchange rates for financial decision problems. Thus, this methodology will allow users who have one or more scenario generation methods implemented to test how well they perform for a given repeated optimization problem without or before explicitly solving the computationally intensive SP. According to the Kaut and Wallace the first two requirements of a good scenario set are estimating the optimality gap and stability. Stability requirement is defined as, if several scenario sets are generated with the same input and same scenario generation approach, and the optimization problem is solved with these scenario sets, approximately the same optimal objective function value should be obtained. This means the optimality gap should be negligible and approximately the same for each scenario set generated by the same method. With this approach, we still must solve the SP to perform scenario assessment. Our aim is to compare scenario sets that are obtained by various scenario generation methods or to assess a prospective scenario generation method without solving the actual SP problem repeatedly. With the proposed approach, quick assessment of scenarios that predicts their performance before solving the SP will decrease the computational burden and time requirements of scenario evaluation.

In a recent study on scenario assessment in the context of power system planning, Pinson and Girard [5] discussed statistical metrics for assessing the quality of wind power scenarios, however; they did not examine the scenarios' performance in a SP problem. Sari et al. investigated the use of statistical evaluation metrics for assessing wind power scenarios for use in SUC [1]. However, some of the metrics proposed in that study were specific to the SUC problem. In this study we aim to propose an approach that is general enough to be applied to any two-stage stochastic program where the observational data on the uncertain parameters are available for some historical time period.

When solving SP models, run time requirements might force the use of a smaller number of scenarios. Thus, scenario reduction techniques are frequently used to trim down the number of scenarios included. It is crucial for a scenario reduction approach to keep most of the stochastic information embedded in scenarios. Scenario reduction concepts are discussed extensively by Heitsch & Römisch [6] and Dupacova et al. [7]. A stability approach led to the forward selection and backward reduction heuristics, which are commonly used scenario reduction techniques for SP problems [6, 7]. A reduced number of scenarios that best retains the essential features of a given original scenario set according to a probability metric can be obtained with these algorithms. The upper bound on the distance between optimal value of the problem with the reduced scenario set and the optimal value of the solution to the original problem is minimized if the scenario sets are sufficiently close in terms of the probability distance [7]. The most common probability distance used for stochastic optimization problems is the Kantorovich distance (mass transportation distance). The minimization problem to calculate the Kantorovich distance between two probability distributions is referred to as Monge-Kantorovich mass transportation problem and the details of the problem is studied in [2]. For two-stage problems, Kantorovich distances are used to derive several heuristics for scenario reduction, including forward selection and backward reduction. According to forward selection heuristic, the scenario that minimizes the Kantorovich distance between the reduced and original sets is selected iteratively in [6, 8]. For our proposed methodology we employ a reliability metric that is motivated by mass transportation distances for evaluating the scenarios [1]. In traditional approaches to scenario reduction, these algorithms are applied directly to scenario distributions. Reduction is based on the norm of the difference between pairs of random vectors. The effect of scenarios on optimal solutions is not addressed directly.

Another approach for using a limited number of scenarios is the sample average approximation method, where quality of solution is assessed by computing statistical bounds on the optimality gaps [9]. Our focus is on scenario generation methods that approximate stochastic processes rather than sampling from their possible realizations.

In recent power system planning studies, researchers have devised scenario reduction techniques based on the optimal objective function values of single-scenario problems. In these approaches, forward selection or backward reduction heuristics are modified to account for the impact of each single scenario realization on the objective function of the stochastic problem. Some numerical evidence indicates that the new scenario reduction procedures outperform the traditional ones [10, 11]. Morales et al. applied such a scenario reduction technique and compared the results with the existing forward selection results [11]. The reduced set of scenarios that is obtained by the proposed technique gives more similar results to those of the original set of scenarios in the SP than does the reduced set of scenarios that is obtained by the existing scenario reduction approach. The superiority of the new approach is illustrated by two different two-stage stochastic problems in the electricity market solved by the producer and the retailer. A scenario reduction method of Bruninx et al. [12] depends on the objective value of the single scenario equivalent of the stochastic problem. Their approach is similar to that of Morales et al.; however, Bruninx et al. do not fix the first stage decision variables whereas Morales et al. compute the cost of the one-scenario equivalent problem with first-stage decision variables fixed to values obtained by solving the expected value problem. Similarly, a heuristic scenario reduction method that selects scenarios based on their cost and reliability impacts is presented by Feng & Ryan [10]. In SUC, they found that fewer load imbalances result from the proposed reduction technique, which clusters scenarios according to their impact on solutions and then applies the fast forward selection heuristic. These approaches are reminiscent of importance sampling, which inspired a scenario selection procedure developed by Papavasiliou et al [13]. They select uncertain scenarios for SUC on the basis of their likelihood of occurrence and the severity of their impact on operating costs.

We propose a scenario assessment approach inspired by the recent scenario reduction techniques. Our proposed methodology also accounts for the impact of each single scenario realization and the evaluation according to the reliability metric is summarized in the form of MTD rank histograms. We demonstrate the proposed methodology in the context of unit commitment and server location problems. Because the interest in stochastic optimization-based unit commitment grown rapidly in the past several years due to deepening penetration of renewable energy [14-18], we focus on the SUC problem as our case study.

4.3 Scenario quality assessment by historical simulation

A good scenario generation method (SGM) should result in low costs in historical simulation over a long sequence of instances. The formalization of scenario quality assessment by historical simulation and the proposed approach will be explained step by step on an abstact form of the general two-stage SP with fixed recourse. A formulation of a generic two-stage SP is [20]:

$$(\mathbb{P}) \qquad \min_{x} \quad c^T x + \mathrm{E}_{\xi} Q\left(x, \xi(\omega)\right) \qquad (1)$$

$$\text{s.t.} \quad Ax = b, \qquad (2)$$

$$x \in X, \qquad (3)$$

where

$$Q\left(x, \xi(\omega)\right) = \min_{y} \ \left\{ q(\omega)^T y \mid \mathrm{T}(\omega) x + Wy = h(\omega),\ y \in Y \right\} \qquad (4)$$

The first stage decisions, $x,$ must be taken without full information on random events, $\omega \in \Omega.$ The random vector, $\xi,$ comprises the parameters of the second-stage problem, $\xi(\omega) = \left(q(\omega), h(\omega), \mathrm{T}(\omega)\right)$. The second stage decisions, denoted by y, are taken after a realization of $\xi(\omega)$ becomes known. Either of the feasible sets X and Y may include integer restrictions. At the first stage, optimization is achieved by minimizing the cost of first-stage decisions, $c^T x$, plus the expected cost of optimal second-stage decisions. When the uncertain data are revealed, optimal second-stage costs are obtained by minimizing $Q\left(x,\xi\left(\omega\right)\right)$ with respect to y. We restrict attention to fixed recourse models with deterministic W.

Using a set of historical instances, the ideal way to assess scenario quality is as follows: For each instance in the historical simulation, generate scenarios using historical data available up to that time and employ them in the SP problem. Simulate the implementation of the first-stage decisions, followed by the second-stage decisions according to the observationl data for that instance. The historical instance of $\mathbb{P}$ for $d\in\{1,2,...,D\}$ is:

$$
\begin{aligned}
&\left(\mathbb{P}^d\right) \qquad && \min_x \quad \left(c^d\right)^T x+\mathrm{E}_{\xi^d} Q\left(x,\xi^d\left(\omega\right)\right) && (5)\\
& && \text{s.t.} \quad A^d x=b^d, && (6)\\
& && \qquad x\in X && (7)
\end{aligned}
$$

We assume we have a corresponding set of historical observations $\left\{\left(q_o^d,h_o^d,T_o^d\right)\right\}_{d=1}^D$ and scenario sets $\left\{\left(q_s^{dk},h_s^{dk},T_s^{dk}\right),\ s\in \mathrm{S}_d\right\}_{d=1}^D$ generated by SGM k , under assessment, along with the set of corresponding probabilities, $\left\{p_s^{dk},\ s\in\mathrm{S}_{dk}\right\}_{d=1}^D$ where $0\le p_s^{dk}\le 1,\ \sum_{s\in S_{dk}} p_s^{dk}=1$ and $k=1,2,...,K$. This produces a collection of extensive forms generated by SGM k :

$$
\begin{aligned}
&\left(\mathbb{P}^{dk}\right) \qquad && \min_x \quad \left(c^d\right)^T x+\sum_{s\in S_{dk}} p_s^{dk} Q^{dk}\left(x,s\right) && (8)\\
& && \text{s.t.} \quad A^d x=b^d, && (9)\\
& && \qquad x\in X, && (10)
\end{aligned}
$$

where

$$
Q^{dk}\left(x,s\right)=\min_y \ \left\{q_s^{dk} y \mid W^d y=h_s^{dk}-T_s^{dk}x,\ y\in Y\right\} \qquad (11)
$$

Let x^{dk} be an optimal solution to $\mathbb{P}^{dk}$. For each $k=1,2,...,K$, we apply the historical simulation as follows:

```
for each $d \in \{1,2,...,D\}$,
        solve $\mathbb{P}^{dk}$ for $x^{dk}$
        solve $Q^{do}\left(x^{dk}\right) = \min_{y} \left\{q_o^d y \mid W^d y = h_o^d - T_o^d x^{dk},\ y \in Y\right\}$    $(12)$
        set $z_o^{dk} = \left(c^d\right)^T x^{dk} + Q^{do}\left(x^{dk}\right)$    $(13)$
compute $c^k = \frac{1}{|D|} \sum_{d \in D} z_o^{dk}$    $(14)$
```

We claim that SGM i has a better quality than SGM j if $c^i < c^j$. Solving $\mathbb{P}^{dk}$ for each $d \in \{1,2,...,D\}$ and $k \in \{1,2,...,K\}$ is hard because of the challenging computational complexity of the SP. Thus, we replace this process with a computationally easier method.

4.4 Proposed generic approach for assessment of scenario generation methods

The proposed scenario generation assessment approach accounts for the impact of each single scenario realization on the optimal cost. A rank histogram is employed to assess the reliability of scenario sets, where the ranks are computed based on the MTD [19]. Although, in general, the MTD is found by solving a linear program, in our application it is the minimum cost of transporting the probability from the group to the individual and can be found in a single step. The notations below relate to computing the MTD and constructing the MTD rank histogram:

u : characteristic of scenario or observation

u_s^d : the value of u in scenario s for instance d

u_o^d : the observed value of u in instance d

$\delta(u',u)$: distance metric

Given observations, u_o^d, and scenario-probability pairs, $V^d = \left\{\left(u_s^d, p_s^d\right)\right\}_{s=1}^{|S_d|}$, for a given set of instances along with a pre-rank function $f\left(u',V;\delta\right) = \sum_{(u,p)\in V} \delta(u',u)p,$ the MTD rank histogram is constructed as follows [1]:

For each $d = 1, 2, ..., D$:

1. Find l_o^d , the distance from the scenarios to the observation:

$$l_o^d \equiv f\left(u_o^d, V^d; \delta\right) \qquad (15)$$

1- For each $s = 1, ..., |S_d|$ compute l_s^d as the distance from the set $S_d \cup \{o\} \setminus s$ to the scenario s where probability p_s^d is assigned to u_o^d , and compute

$$l_s^d \equiv f\left(u_s^d, V^d \setminus \left(u_s^d, p_s^d\right) \cup \left\{\left(u_o^d, p_s^d\right)\right\}\right) \qquad (16)$$

2- Find the rank of l_o^d, denoted r^d, when $\left\{l_o^d\right\} \cup \left\{l_s^d\right\}_{s=1}^{|S_d|}$ are ordered from largest to smallest.

Construct the histogram of $\left\{r^d\right\}_{d=1}^{D}$.

The MTD rank histogram is able to distinguish between sets of scenarios that are more or less reliable according to their bias, variability and autocorrelation when applied to scenarios directly. MTD rank histograms display a downward trend from left to right for an under-dispersed ensemble of scenarios and an upward trend for an over-dispersed ensemble. Bias overpopulates the small ranks similarly as under-dispersion. For scenarios with a higher (lower) autocorrelation level than the observation, a sloping downward (upward) trend is observed. A hill-shaped MTD rank histogram is observed for scenarios with heterogeneous autocorrelation levels. Flat histograms result when the scenarios are reliable [1]. While in [1] the distances among realizations are measured directly, in this paper we modify the definition of u and its metric, δ.

In our proposed method, distances among scenarios and the observed value are measured by fixing first-stage decisions to a common value and computing the differences among second-stage objective values obtained by solving the single-scenario deterministic sub-problems of the SP problem. Thus, the impact of each scenario in the SP problem is taken into consideration.

Our generic approach for scenario generation method assessment is explained as follows: We have observational data $\left\{\left(q_o^d, h_o^d, T_o^d\right)\right\}_{d=1}^{D}$ and scenario sets $\left\{\left(q_s^d, h_s^d, T_s^d\right), s \in \mathrm{S}_d\right\}_{d=1}^{D}$ generated by the method under assessment. To compute distances among scenarios and the observation for each instance d, we solve a single-scenario problem and obtain the optimal first-stage decision variables; i.e., a candidate solution. Then, the second-stage problem is solved for each scenario as well as the observation with the first-stage decision variables fixed to the candidate solution and the second-stage cost is recorded. The distances among the second stage costs are used as the function δ to construct the MTD rank histogram. The steps are formalized as follows:

For each $d \in \{1, ..., D\}$,

Step 1: Solve a single-scenario (deterministic) version of the SP problem with parameters$\left(\hat{q}^d, \hat{h}^d, \hat{T}^d\right)$.

$$\begin{aligned} \left(\hat{x}^d, \hat{y}^{*d}\right) = \underset{x,y}{\arg\min} \quad & c^T x + \hat{q}^d y && (17) \\ \text{s.t.} \quad & Ax = b, && (18) \\ & \hat{T}^d x + Wy = \hat{h}^d && (19) \\ & x, y \geq 0 && (20) \end{aligned}$$

Step 2: For each $g \in G_d = S_d \cup \{o\}$, solve a single-scenario version of the second stage of the SP by fixing the optimal first stage decision variables, $x = \hat{x}^d$, that are obtained from *Step 1*.

$$\begin{aligned} u_g^d = \min_y \quad & q_g^d y && (21) \\ \text{s.t.} \quad & Wy = h_g^d - T_g^d \hat{x}^d && (22) \\ & y \geq 0 && (23) \end{aligned}$$

Step 3: Construct the MTD rank histogram using u_o^d, $V^d = \left\{\left(u_s^d, p_s^d\right)\right\}_{s=1}^{|S_d|}$ and $\delta\left(u', u\right) = \left|u' - u\right|$.

Variants of assessment approach differ in how we define the single-scenario problem (17) - (20) and are explained in detail in Sections 4.6 and 4.7. Figure 4.1 represents the generic scenario assessment for reference in explaining the variants of the approach. Note that, only input and output parameters differ.

For each $d \in \{1,...,D\}$,

$$
\begin{array}{lll}
 & \text{input} & \text{output} \\
\left\{\begin{array}{l} STEP\ 1:\ \text{solve } (17)-(20) \\ STEP\ 2:\ \text{for each g} \in G_d \\ \qquad\qquad \text{solve } (21)-(23) \end{array}\right. & \begin{array}{l} \left(\hat{q}^d, \hat{h}^d,\ \hat{T}^d\right) \\ \mathrm{G}_d = S_d \cup \{o\} \\ x = \hat{x}^d \end{array} & \left.\begin{array}{l} \hat{x}^d \\ \\ u_g^d \end{array}\right\} \\
STEP\ 3:\ \text{construct MTD} & \left\{u_o^d, V^d\right\}_{d=1}^{D}, \delta(u',u) & \text{MTD rank histogram}
\end{array}
$$

Figure 4.1 Generic scenario assessment

4.5 Motivating models

As motivating examples, we consider two challenging stochastic mixed integer programming problems. The stochastic server location problem (SSLP) and the SUC problem are briefly described in this section.

The SSLP is to choose locations of servers from potential locations and allocate clients to the chosen servers to maximize the total expected net revenue subject to the given constraints [21]. Network design for electric power, internet services, telecommunications, and water distribution are some applications of the SSLP. This problem is formulated as a two-stage SP model. Binary first stage decisions determine whether or not to invest in a server at each of the potential locations. Second stage decisions, which are also binary, assign clients to each server. There are constraints on the total number of servers that can be installed and the server capacity. Moreover, each available client can be served by at most one server. The uncertainty occurs in the availability of clients. Thus, the scenarios can be represented as binary vectors where a value of 1 denotes that the corresponding client materializes.

The first-stage cost, which is the investment cost of server siting, is denoted by $c^T x$ in (1). The expected second stage cost, as the negative of the revenue obtained by serving material customers is denoted by $\mathrm{E}_{\xi} Q\left(x, \xi(\omega)\right)$ in (1). The constraint on the total number of servers that can be installed is expressed by (2). Binary restrictions on the first-stage decision variables are expressed by (3). Unserved demand due to the limitations of server capacity (which results in a loss of revenue), the requirement that each available client is served by at most one server, and binary restrictions on the second-stage decision variables are summarized in the feasible region described by (4).

Unit commitment is an important short-term planning problem for electric power generation in which a commitment schedule is identified for each thermal generating unit over a planned time period [22]. In our application, we consider a two-stage SUC formulation where the binary commitment decisions are made in the first stage and the dispatch decisions for the committed units are made in the second stage [23].

The objective, represented by equation (1), is to minimize the total cost which includes the start-up and shut-down costs in the first stage and the expected generation costs along with the heavy penalties on load mismatch in the second stage subject to the operational constraints considering all scenarios. In our application, scenarios represent probabilistic time series for wind energy. Operational constraints include minimum up and down time constraints represented by equation (2), along with energy balance, ramp rate limits and generation level limitations that are summarized in the feasible region described by (4). The uncertain parameters appear in the net load; i.e., the load less the wind energy, on the right-hand-side of the energy balance constraint for each time period. If the total amount of dispatched energy from the committed units is less than the net load then a shortage occurs or, inversely, if it is greater than an excess occurs.

4.6 Expected value based scenario assessment

In recent studies, scenario reduction techniques depend on the objective function of a single-scenario problem, where the impact of each scenario on the SP problem is taken into account. Similar to Morales' approach [11], we judge a scenario set based on the results of each single scenario by fixing the first-stage variables to the values that are obtained from solving the expected-value problem, which constitutes a first approximation of the optimal values of decision variables.

We solve a single scenario version of the SP problem with the expected value (EV) of the scenarios, $\left(\bar{q}^d, \bar{h}^d, \bar{T}^d\right)$. Optimal values of the first-stage decision variables, $\bar{x}^d$, are obtained. For each scenario and observation, we then solve a single-scenario version of the second stage of the SP after fixing the first stage decision variables to $\bar{x}^d$. Second stage costs for each scenario, $\bar{u}_s^d$, and the observation, $\bar{u}_o^d$, are obtained. This approach is summarized in Figure 4.2.

	input	output
$STEP\ 1:$	$\left(\bar{q}^d, \bar{h}^d, \bar{T}^d\right)$	$\bar{x}^d$
$STEP\ 2:$	$\mathrm{G}_d = S_d \cup \{o\}$	
	$x = \bar{x}^d$	$\bar{u}_g^d$
$STEP\ 3:$	$\left\{\bar{u}_o^d, V^d\right\}_{d=1}^{D}, \delta\left(u', u\right)$	MTD rank histogram

Figure 4.2 EV-based scenario assessment

The MTD rank histogram is expected to be flat if the scenarios are of high quality. A simulation study that shows the results of this approach when applied to SSLP is described next. The aim of this simulation study is to show the results of EV-based scenario assessment when scenarios have defects and when scenarios are reliable.

In the test instance used for the EV-based scenario verification simulation study, we have 5 locations and only one server can be installed at each location. Scenario-independent instance data are obtained from [24], which specifies the set of potential server and customer locations, server capacities, installation costs, and revenues. The number of potential clients in this instance is 50. Scenario-dependent instance data additionally specifies the set of customers that are actually realized in that specific scenario. Each simulation study consists of 1000 instances and each scenario set consist of 10 scenarios. For the first simulation study we generated the scenario and observational data for client availability from independent Bernoulli distributions with a common parameter, the probability a client is available. For observational data, the probability of availability of each client is $p_{obs} = 0.5.$ To test whether the proposed approach detects bias, we set p_{scen} to 0.1, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.9. Figure 4.3 shows the MTD rank histograms.

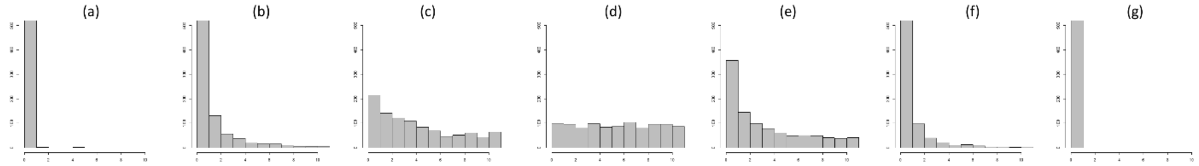


Figure 4.3 MTD rank histograms obtained by EV-based scenario assessment for various values of parameter $p_{scen} = (a)\ 0.1\ (b)\ 0.3\ (c)\ 0.4\ (d)\ 0.5\ (e)\ 0.6\ (f)\ 0.7\ (g)\ 0.9$ when $p_{obs} = 0.5$.

A downward trend appears in all the MTD rank histograms except in panel (d) where $p_{obs} = p_{scen}$. The magnitude of the slope of the MTD rank histograms increases with the difference between p_{scen} and p_{obs} . This is due to the increasing bias in the second stage cost results. When the parameters are equal for the scenarios and observational data we observe a flat histogram (c), which is an indicator of high quality scenarios.

Figure 4.4 shows the results of simulating the proposed approach when there is no bias ($p_{obs} = p_{scen} = 0.5$) but there are correlation inconsistencies between observations and scenarios. We generated correlated binary variables by using the exchangeable correlation structure method of [25]. For observational data the pairwise correlation of availability of each client is $\rho_{obs} = 0.01$, whereas for scenario data, we set $\rho_{scen} = 0.0001,\ 0.0025,\ 0.01,\ 0.04,\ 0.36,$ and $0.81,$ respectively.

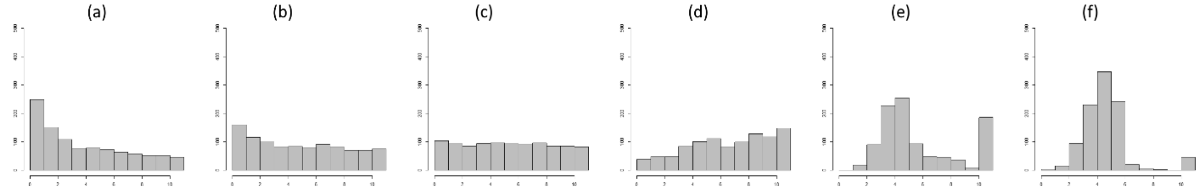


Figure 4.4 MTD rank histograms obtained by EV-based scenario assessment for various values of parameter $\rho_{scen} = (\mathrm{a})\ 0.0001\ (\mathrm{b})\ 0.0025\ (\mathrm{c})\ 0.01\ (\mathrm{d})\ 0.04\ (\mathrm{e})\ 0.36\ (\mathrm{f})\ 0.81$when $\rho_{obs} = 0.01$.

When scenarios and the observation are generated with the same probability and correlation, the MTD rank histogram appears to be flat in panel (c), indicating the scenarios are of good quality. As we increase the correlation of scenarios to $.2^2$ we observe an upward trend in (d) which means the second stage costs of the scenarios are over-dispersed. As we continue to increase the difference of correlation between scenario and observation data, hill-shaped rank histograms occur. These indicate that the range of the MTDs among the costs of scenarios is wide, so the MTD from scenarios to the observation falls in the middle frequently. The hill shape in rank histograms becomes more pronounced as the difference between correlations is increased as shown in (e) and (f). When scenarios have lower correlation than the observation as in panels (a) and (b), we see a downward trend that is steeper when the difference is greater.

EV-based scenario assessment is a very useful approach for assessing scenario quality. With this approach we can verify a scenario set to be of high quality for the related SP problem and compare the variants of scenario generation approach. Moreover, the results are easy to interpret. For good quality scenarios we expect a flat rank histogram because the second stage cost of observation should be indistinguishable among the second stage cost of scenarios.

Thus, the uniformity of the histograms can be tested with a goodness-of-fit test. However, this approach may not differentiate among distinct scenario generation approaches. To make a better comparison on the distinct scenario generation methods, we should evaluate them on the same basis. Thus, using the same first-stage decisions for both scenario sets, we may make a better comparison among distinct scenario sets. With the perfect information based scenario assessment explained in Section 4.7, first-stage decision variables are obtained through the hindsight available in a historical simulation.

4.7 Perfect information based scenario assessment

We solve single scenario version of the SP problem with the perfect information (PI) $\left(q_o^d, h_o^d, T_o^d\right)$ to obtain optimal values of the first- and second-stage decision variables, $x_o^d, \breve{u}_o^d$. We fix the first stage decision variables to x_o^d and, for each scenario and observation, solve a single-scenario version of the second stage of the SP problem. Second stage costs for each scenario, $\breve{u}_s^d$, are obtained. This approach is summarized in Figure 4.5.

	input	output
STEP 1:	$\left(q_o^d, h_o^d, T_o^d\right)$	$x_o^d, \breve{u}_o^d$
STEP 2:	$G_d = S_d$	
	$x = x_o^d$	$\breve{u}_g^d$
STEP 3:	$\left\{\breve{u}_o^d, V^d\right\}_{d=1}^{D}, \delta\left(u', u\right)$	MTD rank histogram

Figure 4.5 PI-based scenario assessment

4.7.1 Simulation study on server location problem

The results of PI-based scenario assessment when applied to our server location instance are shown in Figures 4.6 and 4.7. We used the same sets of scenarios and observations as for the simulation study in Section 4.6. The results of PI-based scenario verification are similar to the results of EV-based scenario assessment. When scenarios and observations are drawn from the same distribution, we observe flat MTD rank histograms. When there is bias or correlation inconsistencies in scenarios, downward trends or hill shapes are observed.

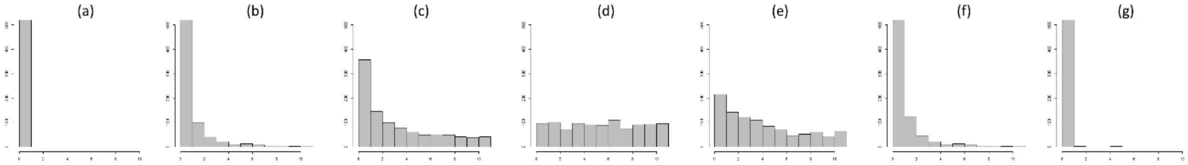


Figure 4.6 MTD rank histograms obtained by PI-based scenario assessment for various values of parameter $p_{scen} = (\mathrm{a})\ 0.1 \quad (\mathrm{b})\ 0.3\,(\mathrm{c})\ 0.4 \quad (\mathrm{d})\ 0.5 \quad (\mathrm{e})\ 0.6 \quad (\mathrm{f})\ 0.7 \quad (\mathrm{g})\ 0.9$ when $p_{obs} = 0.5$.

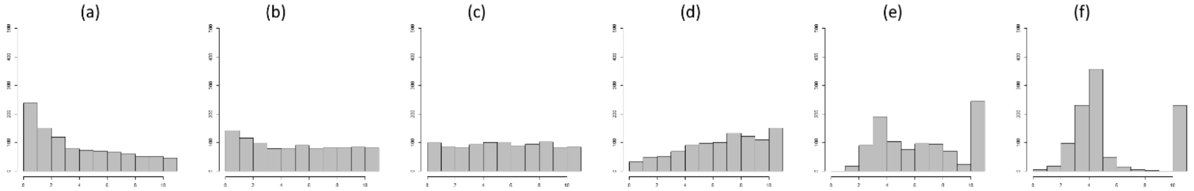


Figure 4.7 MTD rank histograms obtained by PI-based scenario assessment for various values of parameter $\rho_{scen} = (\mathrm{a})\ 0.0001 \quad (\mathrm{b})\ 0.0025 \quad (\mathrm{c})\ 0.01 \quad (\mathrm{d})\ 0.04 \quad (\mathrm{e})\ 0.36 \quad (\mathrm{f})\ 0.81$ when $\rho_{obs} = 0.01.$

When the optimal objective value of the SP is sensitive to the optimal solution, the resulting MTD rank histogram may not expected to be flat. Because we evaluate the scenarios' results by fixing the first stage decisions to the optimal first stage decision variables obtained through the observational data, the second stage costs of scenarios are expected to be almost always higher than the costs of observational data. Moreover, the range of the distances among the second stage costs of scenarios might be large. We would expect a downward trend in the rank histogram if the bias in second stage costs of scenarios is dominant or a hill-shaped rank histogram if the width of the range of the distances among second stage costs is dominant even though the scenarios are of high quality. This phenomenon is illustrated by the SUC problem as explored in a simulation study in Section 4.7.2.

4.7.2 Simulation study on unit commitment problem

Simulation studies that show the results of PI-based scenario assessment when applied to SUC are described. The aim of these simulation studies are to show the results of this approach when scenarios have defects (bias, under/over-dispersion, bias that is hidden by variation, autocorrelation inconsistencies) and when scenarios are reliable. The results are shown in Figures 4.8-4.13. For the simulation studies, we generated simulated wind scenarios and observations from AR(1) distribution controlling the parameters of mean $\left(\mu_{obs} \text{ and } \mu_{scen}\right)$, standard deviation $\left(\sigma_{obs} \text{ and } \sigma_{scen}\right)$ and autocorrelation $\left(\rho_{obs} \text{ and } \rho_{scen}\right)$. Each panel represents 1000 instances, with 10 scenarios for instance consisting of 24 hourly values.

Figure 4.8 shows the results of reliable scenarios, where the mean, variation, and autocorrelation of scenarios are equal to those of the observations. We use the same set of observations for simulations shown in Figures 4.10-4.13.

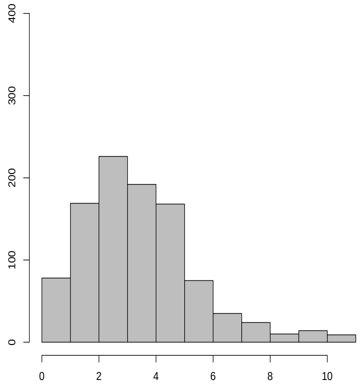


Figure 4.8 MTD rank histograms obtained by PI-based scenario assessment for $\mu_{obs} = \mu_{scen} = 2500,\ \sigma_{obs} = \sigma_{scen} = 100,$ and $\rho_{obs} = \rho_{scen}.70$.

The MTD rank histogram in Figure 4.8 is not flat for reliable scenarios when applied to unit commitment problem. The middle ranks are over populated because the second stage cost of observational data is almost always lower than the second stage costs of scenarios and the range of differences among cost of scenarios is large. This causes the second stage costs of observational data to fall in the middle ranks. As a result, we observe a hill-shaped rank histogram for reliable scenarios.

For reliable scenarios, we may observe downward trends, hill-shapes, or right skewed hill-shapes as we change the parameter settings. Some possible outcomes are presented in Figure 4.9. It is inevitable that the second stage cost of observational data is almost always lower than the second stage costs of scenarios which may cause bias in second stage costs of scenarios. If the bias is pronounced, we see downward slope in the rank histograms. Moreover, the range of the distances among the second stage costs of scenarios might be large which cause a hill-shaped rank histogram. If the width of the range of the distances among second stage costs is much dominant, we see hill-shaped rank histograms. Some combinations of different parameters can cause hill-shaped rank histograms that is right-skewed.

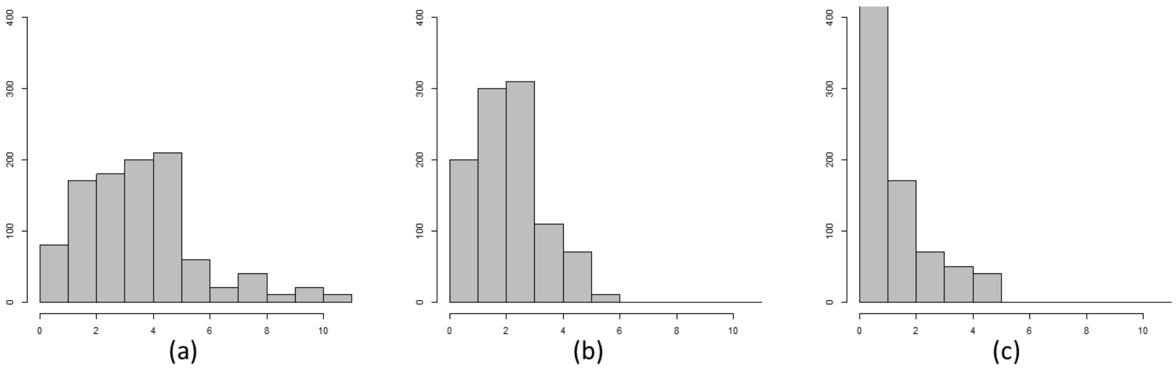


Figure 4.9 Possible MTD rank histograms obtained by PI-based scenario assessment when we try different combinations of parameters while keeping $\mu_{obs} = \mu_{scen}$, $\sigma_{obs} = \sigma_{scen}$, and $\rho_{obs} = \rho_{scen}$ with (a) $\left(s_{obs} = 70, m_{obs} = 2500, r_{obs} = .75\right)$ (b) $\left(s_{obs} = 300, m_{obs} = 1500, r_{obs} = .50\right)$ and (c) $\left(\sigma_{obs} = 800, \mu_{obs} = 800, \rho_{obs} = .50\right)$.

Figure 4.10 shows the results of testing PI-based scenario assessment when applied to unit commitment problem with biased scenarios. We keep $\sigma_{obs}/\sigma_{scen} = 1, \rho_{obs} = \rho_{scen}$ and test various combinations of parameters μ_{obs} and μ_{scen}.

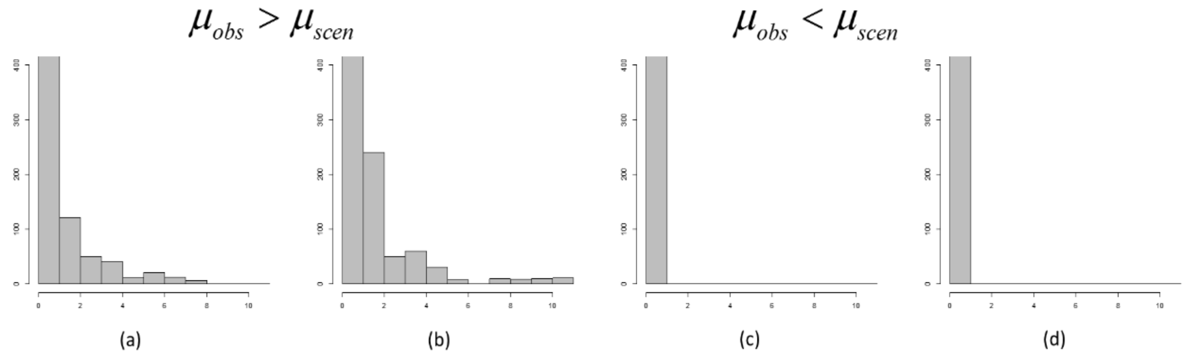


Figure 4.10 MTD rank histograms obtained by PI-based scenario assessment when applied to SUC for $\mu_{obs} = 2500$ and $\mu_{scen} = (\text{a})\ 1500\ (\text{b})\ 2000\ (\text{c})\ 2750\ (\text{d})\ 3000$.

The MTD rank histograms (a) and (b) are constructed by scenarios with lower mean than the observation data ($\mu_{obs} > \mu_{scen}$). This caused a high bias in the resulting second stage costs due to excess penalties. Thus, we observe a downward trend. The amount of bias is higher in (a) than (b), thus the skewness of rank histogram is much pronounced in (a). Similarly, the rank histograms (c) and (d) are constructed by biased scenarios, however, the mean of the scenarios is higher than the observed wind power ($\mu_{obs} < \mu_{scen}$). The amount of bias in (a) and (b) are higher than the amount of bias in (c) and (d), but the directions of bias are opposite. As can be seen, only rank 1 occurs in (c) and (d). This is because of the increased bias in the second stage costs because of the direction of the bias in parameter values. When we predict wind scenarios to be higher than the observation, the net load cannot be satisfied and this would cause high second stage costs due to high shortage penalties. The dramatic effect of scenarios is indicated by the rank histograms with only rank 1. All of these cases cause MTD rank histograms to have a downward trend; however, the magnitude of the slope differs according to the bias direction and bias amount. As the impact of misestimation increases in the associated SP, the slope of the rank histogram increases.

Figure 4.11 shows the results of testing PI-based scenario assessment when applied to unit commitment problem with under/over-dispersed scenarios. We keep $\mu_{obs} = \mu_{scen}$, $\rho_{obs} = \rho_{scen}$ and test various combinations of parameters σ_{obs} and σ_{scen} .

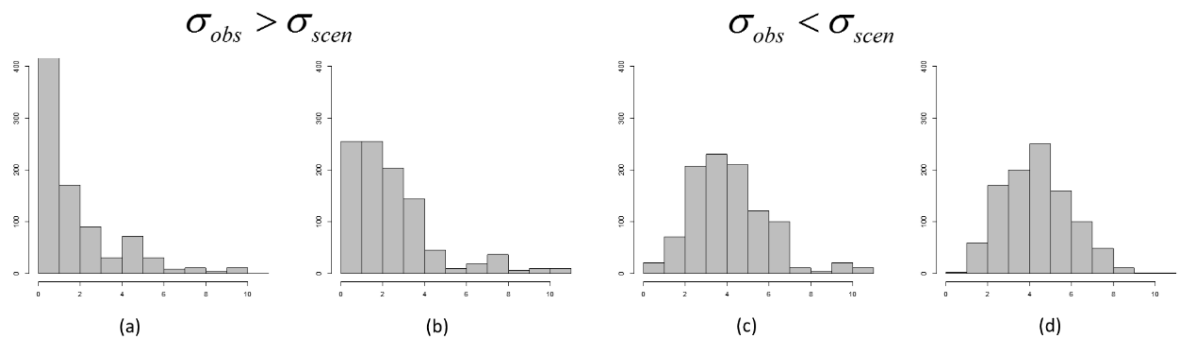


Figure 4.11 MTD rank histograms obtained by PI based scenario assessment when applied to SUC for $\sigma_{obs} = 100$ and $\frac{\sigma_{obs}}{\sigma_{scen}} = (\text{a})\ 4\ (\text{b})\ 2\ (\text{c})\ \frac{1}{2}\ (\text{d})\ \frac{1}{4}$.

The rank histograms (a) and (b) are constructed by under-dispersed scenarios that occurs when the variation in observed data are higher than the scenarios ($\sigma_{obs} > \sigma_{scen}$). Under-dispersion causes the MTD rank histogram to have a downward trend. The slope varies according to the ratio between the variation in scenarios and observation. The rank histograms (c) and (d) are constructed by over-dispersion that occurs when the variation in observed data is lower than the scenarios ($\sigma_{obs} < \sigma_{scen}$). This causes the rank histograms to be hill shaped. As the ratio of the variation increases, the population of the middle ranks increases and the most popular ranks slightly change, as well.

In Figure 4.12, we show the results of testing the scenarios that are both over-dispersed and biased. We use the same bias as in Figure 4.7 (b) and the same over-dispersion as in Figure 4.8 (d).

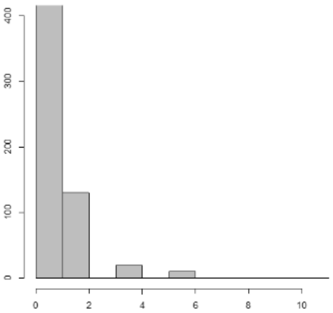


Figure 4.12 MTD rank histogram obtained by PI-based scenario assesment for $\mu_{scen} = 2000$, $\frac{\sigma_{obs}}{\sigma_{scen}} = \frac{1}{4}$, and $\rho_{obs} = \rho_{scen}$.

We observe a downward trend when there is both bias and over-dispersion. As can be seen the magnitude of the slope of the rank histogram is higher than the rank histogram constructed by the scenarios with the same bias but without the over-dispersion. Thus, this will show us the impact of scenarios with bias and over-dispersion would be higher in the SP. In evaluation of raw scenario data, this combination would lead to a misleadingly flat histogram, because high variation in scenarios can compensate for bias in the raw data. Since we are evaluating the impact of scenarios in SP, the rank histogram displays a downward trend with a higher magnitude of slope.

In Figure 4.13 we show the effect of autocorrelation inconsistencies in scenarios. Different autocorrelation levels make a slight change in the resulting rank histograms. When we generate scenarios with heterogeneous autocorrelations which have higher (lower) ρ coefficients than observation, more (less) steep hill-shaped rank histograms are observed.

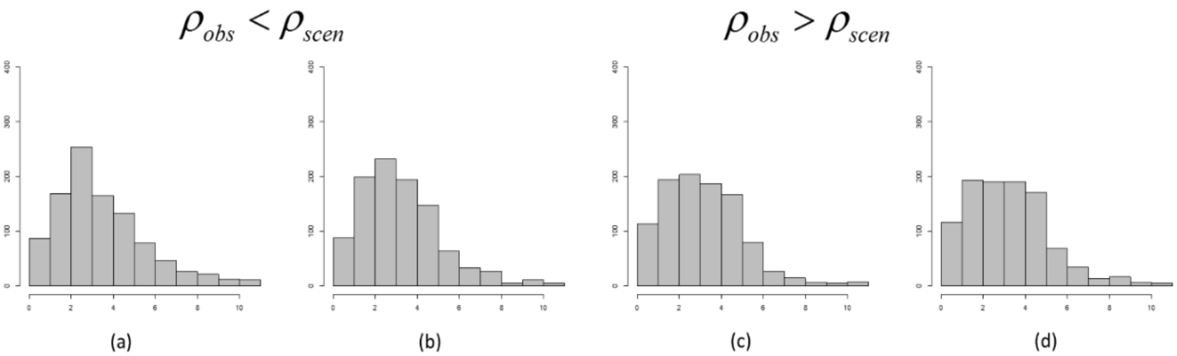


Figure 4.13 MTD rank histograms obtained by PI based scenario assessment for heterogeneous ρ_{scen} when applied to unit commitment problem, where ρ_{scen} consists of (a) $(.70,\ .85, .95)$, (b) $(.70,\ .75, .80)$, (c) $(.70,\ .60, .50)$, (d) $(.60,\ .40, .20)$ where the number of scenarios for the specified values are, 3, 4, and 3, respectively.

Consequently, we cannot expect a flat rank histogram even though the scenarios are of high quality when the optimal objective value of the SP is sensitive to the optimal solution. Thus, we cannot quantify the resulting rank histogram by checking its uniformity for assessing the quality of scenarios.

4.7.3 Evaluation of MTD rank histogram for PI-based scenario assessment

Since a flat rank histogram is not expected, we cannot account for the flatness of the MTD rank histogram of PI-based scenario assessment. Thus, we seek to characterize a better or worse pattern. We conjecture that if the observed data have similar characteristics to the scenario sets, a very similar MTD rank histogram would be constructed by ignoring the observed data and treating one of the scenarios as if it were the observation. Thus, we will assume that a randomly chosen scenario becomes the observed data for the SP. Scenarios producing a rank histogram that is similar to that obtained by PI-based scenario assessment would be expected to perform well in the application.

We solve a single scenario version of the SP problem with a scenario that is chosen randomly$\left(q_{s'}^d, h_{s'}^d, T_{s'}^d\right)$. Optimal values of the first-stage and second-stage decision variables, $x_{s'}^d, \tilde{u}_{s'}^d$ are obtained. For each scenario, we solve a single-scenario version of the second stage of the SP after fixing the first stage decision variables to $x_{s'}^d$. Second stage costs for each scenario, $\tilde{u}_s^d$, are obtained. This approach is summarized in Figure 4.14.

	input	output
$STEP\ 1:$	$\left(q_{s'}^d, h_{s'}^d, T_{s'}^d\right)$	$x_{s'}^d, \tilde{u}_{s'}^d$
$STEP\ 2:$	$G_d = S_d \setminus s'$	
	$x = x_{s'}^d$	$\tilde{u}_g^d$
$STEP\ 3:$	$\left\{V^d\right\}_{d=1}^D, \delta\left(u', u\right)$	MTD rank histogram

Figure 4.14 Random scenario based scenario assessment

The resulting rank histogram should be similar to the rank histogram (rank histogram of PI-based scenario assessment) that is being assessed for high quality scenarios.

When this method is applied using the same set of reliable scenarios as in Figure 8, which is reproduced for convenience in Figure 15(a), we obtain the rank histogram shown in Figure 15(b). The similarity of the two panels confirms that the scenarios are of high quality.

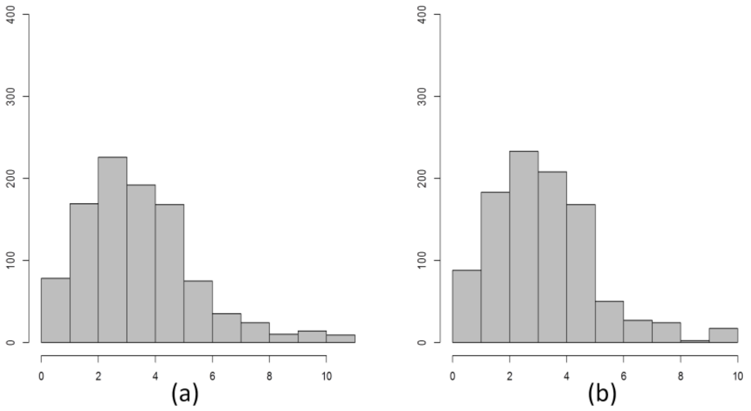


Figure 4.15 MTD rank histogram obtained by a randomly selected scenario to evaluate the MTD rank histogram shown in Fig. 4.8

4.8 Case study of EV-based and PI-based scenario assessment – unit commitment problem

We use the SUC problem as our case study and assume uncertainty occurs only in the wind energy. Thus, scenarios are time series that represent amounts of available wind power in each time period. To generate wind power scenarios, we used the day-ahead wind forecast and observation data from the Bonneville Power Administration from 2012/10/01 to 2013/09/31. The quantile regression with Gaussian copula approach (QR) [26] and epi-spline approximation approach (EPI) [27] are used for scenario generation. We test two variants of each approach labeled as QR, QRnew, EPIwide, and EPInarrow. We obtain the load data from Independent System Operator of New England (ISO-NE). The details of scenario generation methods and how the input data are obtained are documented in [1]. For illustration, in Figure 4.16, we plot the scaled wind power scenarios that are generated for the same day.

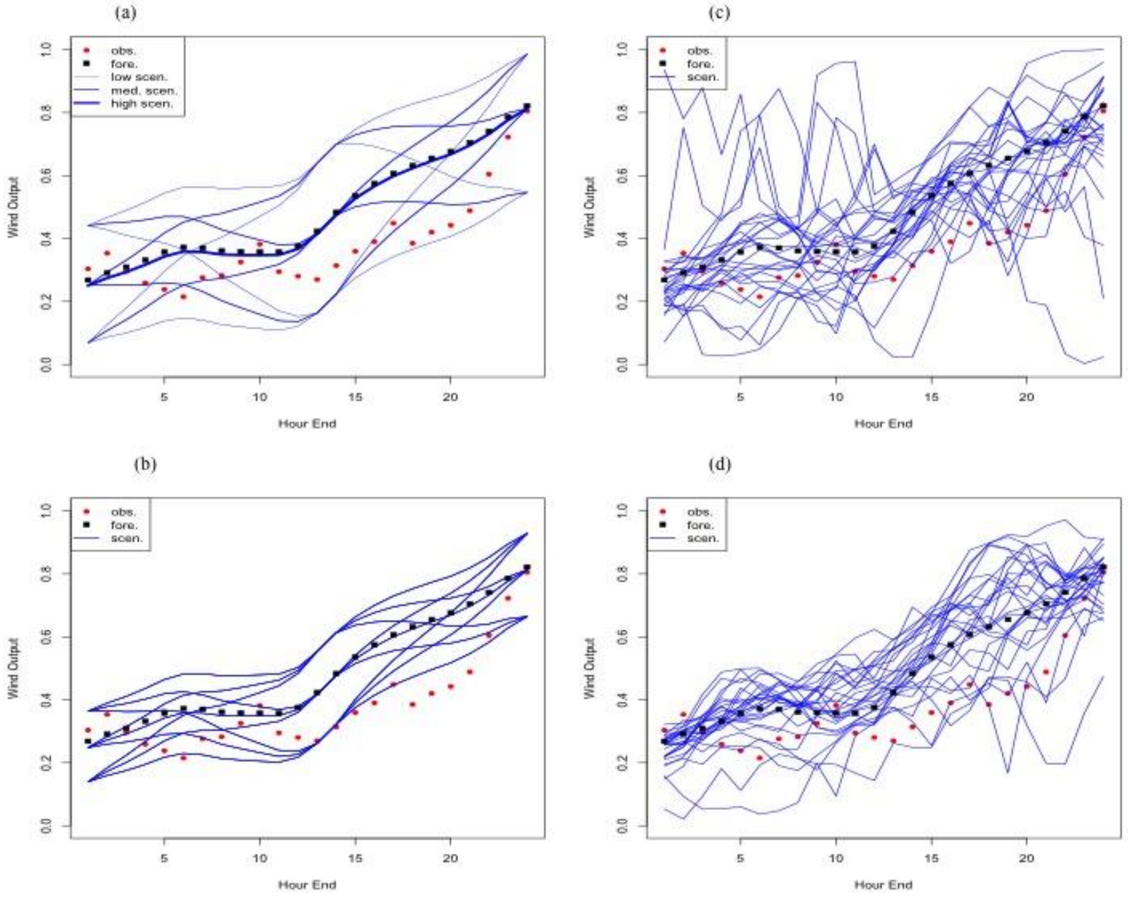


Figure 4.16 Wind power scenarios generated for the same day. (a) EPIwide, (b) EPInarrow, (c) QR, (d) QRnew

In Figure 4.17 we show the results of EV-based scenario assessment when applied to UC problem with EPIwide, EPInarrow, QR, and QRnew scenarios.

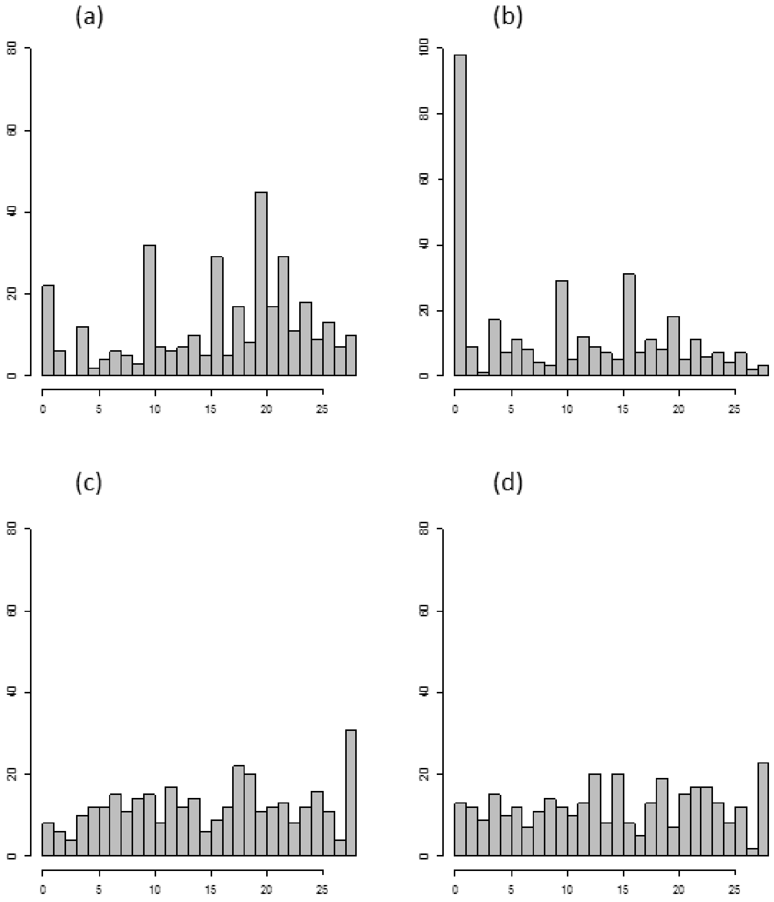


Figure 4.17 MTD rank histograms obtained by EV-based scenario assessment when applied to UC problem with scenarios generated by approaches (a) EPIwide, (b) EPInarrow, (c) QR and (d) QRnew.

The MTD rank histogram constructed by EPInarrow scenarios display a downward trend. The smallest rank is over-populated. This is a result of high bias and/or under-dispersion in the results of second stage costs. There is no obvious trend in the MTD rank histogram constructed based on EPIwide scenarios. According to the MTD rank histograms, we expect to achieve higher cost savings in unit commitment and dispatch problem with EPIwide scenarios than with EPInarrow scenarios. The QR method results in MTD rank histograms with an upward trend because the resulting second stage costs of QR scenarios are over-dispersed. QRnew scenarios result in a flatter rank histogram than QR. Similarly, we expect that solutions obtained with QRnew scenarios will incur lower costs than those found with QR scenarios.

A high quality scenario set should result in a relatively flat histogram, which means the results of observational data are indistinguishable among the results of scenario data. With this approach, we can verify a scenario set to be of high quality for the related SP problem and compare the variants of each scenario generation approach. However, this approach may not differentiate among distinct scenario generation approaches. In order to make a comparison between quantile regression and epi-spline scenario, we need to evaluate them on the same basis.

With the proposed EV-based scenario assessment we can eliminate the scenarios generated by EPInarrow over EPIwide and QR over QRnew. However, when comparing epi-spline and quantile regression scenarios we may not choose the scenario set that is expected to perform better in SUC problem by only relying on the results of MTD rank histogram of EV-based scenario verification. To compare them on the same basis we apply PI-based scenario assessment where we used common 1st stage decision variables. As can be seen in Figure 4.18, a hill-shaped rank histogram is obtained for both EPIwide and QRnew scenarios.

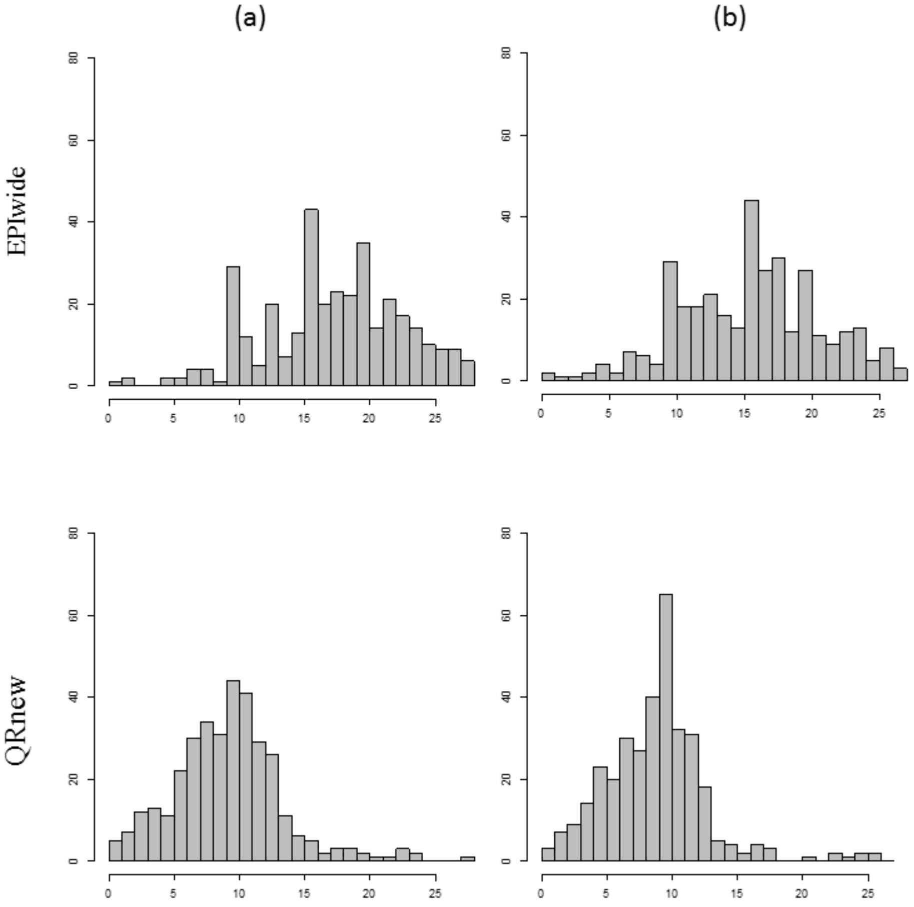


Figure 4.18 MTD rank histograms obtained with scenarios generated by approaches EPIwide and QRnew by (a) PI-based scenario assessment and (b) randomly selected scenario for assessment.

The resulting MTD rank histograms of PI-based scenario assessment and randomly selected scenarios appear similar for both QRnew and EPIwide scenarios. Because there is no a downward trend in the results of PI-based scenario assessment, we assume that scenarios are not biased and/or under-dispersed. However, a less steep hill-shaped rank histogram stands out for PI-based than the randomly selected scenario for QRnew scenarios. As shown in simulation studies a much or less steep rank histogram can be observed when there is autocorrelation inconsistencies in scenarios. Particularly, when there are heterogeneous autocorrelations in scenarios where the correlation levels are lower than the observational data, a less steep rank histogram can be observed (Figure 4.13 (d) and Figure 4.15). A similar case occurred for QRnew scenarios. For EPIwide scenarios there is no obvious difference in the steepness of the hill-shaped rank histogram. The population of some middle ranks slightly changed in Figure 4.11 (c) and (d) compared to Figure 4.15, which might occur for over-dispersed scenarios.

We compute the MTD between empirical distributions of ranks in each bin [28] to make a comparison according to the similarities quantitatively. The MTD between rank histograms of PI-based and random scenario based approaches for EPIwide and QRnew scenarios (which are shown in Figure 4.18) are 1.2976 and 1.8479, respectively. According to this metric, the rank histograms of EPIwide are more similar than those of QRnew.

According to the results represented in the case study, we can conclude that, of the wind power scenario generation methods and variants tested, EPIwide and QRnew perform better in SUC problem as could be predicted by flat MTD rank histograms of EV-based scenario assessment. In order to compare EPIwide and QRnew on the same basis, we applied PI-based scenario assessment. Even though the interpretation of the results of PI-based scenario assessment is not straightforward for unit commitment problem, EPIwide might be expected to perform best as indicated by the similarity of MTD rank histograms of PI-based and randomly selected scenario based scenario assessment.

4.9 Conclusions

High quality scenarios are very important for achieving costs savings by solving SP problems. We proposed EV-based and PI-based scenario assessment approaches aiming to assess the quality of scenarios quickly. We applied them to server location and unit commitment problems with the simulated scenarios to show the results of approaches when scenarios are reliable or unreliable. How to interpret the results of two approaches is explained with the simulation studies. As our case study we use the SUC problem where uncertainty occurs in wind energy. Two different scenario generation methods, along with two variants, are tested with the proposed approaches. We aim to predict each scenario set's unit commitment performance.

With the EV-based scenario assessment, we expect a flat histogram for a high quality scenario set. It is a useful approach when comparing the variants of a scenario generation method. If distinct scenario generation methods are under evaluation, we should apply both the EV-based and PI-based scenario assessments. However, we may not always expect a flat histogram of to result from PI-based scenario assessment. In such cases, we apply generic scenario assessment where we solve the single-scenario problem to obtain first-stage decisions with a randomly chosen scenario. We expect MTD rank histograms of PI-based scenario assessment and randomly selected scenarios to look similar for high quality scenarios. With the proposed approaches, the scenario generation methods that are expected to lead low costs in SP problem can be verified quickly.

REFERENCES

[1] D. Sari, Y. Lee, S. Ryan, and D. Woodruff, "Statistical metrics for assessing the quality of wind power scenarios for stochastic unit commitment," *Wind Energy*, vol. 19, pp. 873-893, May 2016.

[2] S. T. Rachev, *Probability metrics and the stability of stochastic models*. Chichester: Wiley, 1991.

[3] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems*. Berlin: Springer-Verlag 1998.

[4] M. Kaut and S. W. Wallace, "Evaluation of scenario-generation methods for stochastic programming," *Pacific Journal of Optimization*, vol. 3, pp. 257-271, 2007.

[5] P. Pinson and R. Girard, "Evaluating the quality of scenarios of short-term wind power generation," *Applied Energy*, vol. 96, pp. 12-20, 8// 2012.

[6] H. Heitsch and W. Römisch, "Scenario Reduction Algorithms in Stochastic Programming," *Computational Optimization and Applications*, vol. 24, pp. 187-206, 2003.

[7] J. Dupacova, N. Gröwe-Kuska, and W. Römisch, "Scenario reduction in stochastic programming: An approach using probability metrics," *Mathematical Programming*, vol. 95, pp. 493-511, 2003.

[8] H. Heitsch, W. Römisch, "A note on scenario reduction for two-stage stochastic programs," *Oper. Res. Lett.*, vol. 35, pp. 731-738, 2007.

[9] S. Ahmed and A. Shapiro, "The Sample Average Approximation Method for Stochastic Programs with Integer Recourse," Technical Report, ISyE, Georgia Tech2002.

[10] Y. Feng and S. M. Ryan, "Solution sensitivity-based scenario reduction for stochastic unit commitment," *Computational Management Science*, vol. 13, pp. 29-62, 2016.

[11] J. M. Morales, S. Pineda, A. J. Conejo, and M. Carrion, "Scenario Reduction for Futures Market Trading in Electricity Markets," *IEEE Transactions on Power Systems*, vol. 24, pp. 878-888, 2009.

[12] K. Bruninx, E. Delarue, and W. D‘haeseleer, "A practical approach on scenario generation & reduction algorithms for wind power forecast error scenarios," Retrieved from https://www.mech.kuleuven.be/en/tme/research/energy_environment/Pdf/wp2014-15b.pdf2014.

[13] A. Papavasiliou and S. S. Oren, "Multiarea Stochastic Unit Commitment for High Wind Penetration in a Transmission Constrained Network," *Operations Research*, vol. 61, pp. 578-592, 2013.

[14] E. A. Bakirtzis, P. N. Biskas, D. P. Labridis, and A. G. Bakirtzis, "Multiple Time Resolution Unit Commitment for Short-Term Operations Scheduling Under High

Renewable Penetration," *IEEE Transactions on Power Systems,* vol. 29, pp. 149-159, Jan 2014.

[15] K. Bruninx, K. Van den Bergh, E. Delarue, and W. D'haeseleer, "Optimization and Allocation of Spinning Reserves in a Low-Carbon Framework," *IEEE Transactions on Power Systems,* vol. 31, pp. 872-882, 2016.

[16] K. Bruninx, Y. Dvorkin, E. Delarue, H. Pandžić, W. D'haeseleer, and D. S. Kirschen, "Coupling Pumped Hydro Energy Storage With Unit Commitment," *IEEE Transactions on Sustainable Energy,* vol. 7, pp. 786-796, 2016.

[17] H. Wu and M. Shahidehpour, "Stochastic SCUC Solution With Variable Wind Energy Using Constrained Ordinal Optimization," *IEEE Transactions on Sustainable Energy,* vol. 5, pp. 379-388, 2014.

[18] Q. P. P. Zheng, J. H. Wang, and A. L. Liu, "Stochastic Optimization for Unit Commitment-A Review," *IEEE Transactions on Power Systems,* vol. 30, pp. 1913-1924, Jul 2015.

[19] D. Sari and S. Ryan, "MTDrh: Mass Transportation Distance Rank Histogram," ed. https://cran.r-project.org/web/packages/MTDrh/index.html, 2016

[20] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. New York: Springer, 1997.

[21] L. Ntaimo and S. Sen, "The Million-Variable “March” for Stochastic Combinatorial Optimization," *Journal of Global Optimization,* vol. 32, pp. 385-400, 2005.

[22] S. Takriti, J. R. Birge, and E. Long, "A stochastic model for the unit commitment problem," *IEEE Transactions on Power Systems*, vol. 11, pp. 1497-1508, 1996.

[23] Y. H. Feng, I. Rios, S. M. Ryan, K. Spurkel, J. P. Watson, R. J-B Wets, D. L. Woodruff, "Toward scalable stochastic unit commitment. Part 1: load scenario generation," *Energy Systems-Optimization Modeling Simulation and Economic Aspects*, vol. 6, pp. 309-329, Sep 2015.

[24] http://www2.isye.gatech.edu/~sahmed/siplib/sslp/sslp.html *last visited February 10, 2017*.

[25] A. D. Lunn and S. J. Davies, "A note on generating correlated binary variables," *Biometrika*, vol. 85, pp. 487-490, 1998.

[26] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klockl, "From Probabilistic Forecasts to Statistical Scenarios of Short-term Wind Power Production," *Wind Energy*, vol. 12, pp. 51-62, Jan 2009.

[27] I. Rios, R. J-B Wets, and D. L. Woodruff, "Multi-period forecasting and scenario generation with limited data," *Computational Management Science*, vol. 12, pp. 267-295, 2015.

[28] https://cran.r-project.org/web/packages/transport/transport.pdf *last visited March 17, 2017*.

CHAPTER 5: GENERAL CONCLUSION

In this dissertation, we aim to assess the quality of scenarios and scenario generation methods for stochastic programming problems. We focus on the stochastic unit commitment (SUC) and dispatch problem as our case study in chapters 2 and 3. In chapter 4 we extend this work to make it more robust and general enough to be applied to any two-stage stochastic programming problem that is repeatedly solved and for which observational data exist for some historical period.

In chapter 2, we test wind power scenarios for use in stochastic unit commitment for reliability, sharpness, skill and their ability to capture critical characteristics of stochastic processes by examining several verification tools. MST rank histograms are employed to assess the temporal dependence structure of equally likely scenarios. A new MTD rank histogram is developed for assessing the reliability of scenarios that may have different probabilities of occurrence, where reliability is defined as the statistical consistency between scenarios and observation. By simulation studies, it is shown that the MTD rank histogram can distinguish scenario sets that are calibrated according to their bias, variability and autocorrelation. This tool is applied to distinguish between sets of scenarios generated by two very different approaches: quantile regression with Gaussian copula, and epi-spline approximation. Energy scores and event-based Brier scores are applied to compare and contrast multiple attributes of the scenario sets. Energy scores are used to inform on the forecast skill of scenarios for individual lead times. Event-based verification assesses the ability of wind power scenarios to accurately represent ramp up and ramp down events, which can have a large impact on unit commitment and subsequent dispatch costs.

In chapter 3, we examine the relationship between the wind power scenario assessment metrics, which are identified in chapter 2, and the resulting SUC and dispatch costs. A convincing way to evaluate a scenario generation method is to simulate employing the resulting scenarios in SUC while measuring the costs incurred. However, although advanced methods are applied to SUC problems to reduce the computational effort, a simulation study sufficiently thorough to accurately detect meaningful differences among scenario sets over a historical time period remains computationally very demanding. Therefore, we explore the ways to evaluate a set of scenarios or scenario generation method without extensively simulating the SUC procedure. In this manner, we distinguish the effects of scenario sets on the SUC solutions and, without extensively simulating the SUC, choose a scenario generation method that is expected to yield low costs. The scenario sets that might lead to either high no-load costs or high penalty costs due to shortage or excess can be eliminated by the lack of uniformity in a MTD rank histogram. Event-based metrics can help to predict the cost performance of the remaining scenario sets. The limitations of both statistical metrics are also discussed.

In chapter 4, we complete the work by extending the scenario assessment approach to a more general class of two-stage stochastic programming problems. We evaluate second stage costs resulting from the scenarios when the first stage decision variables are fixed to common values in the single-scenario deterministic equivalent problem of the associated two-stage stochastic program. We utilize two novel approaches: expected value (EV) based and perfect information (PI) based scenario assessment. These assessment approaches are not only applicable to a variety of two-stage stochastic problems, but are also expected to approximate the results of simulation-based assessment. Instead of assessing the fit of scenarios to observations directly, we evaluate their impacts on the solutions and resulting costs, which might be a good approximation of the results of employing them in the stochastic program. With the proposed approach, we assess scenarios generation results in the context of a target optimization problem quickly.

The contributions of each chapter are summarized as follows.

In chapter 2:

- The MTD rank histogram is developed to assess the reliability of multi-dimensional scenarios with unequal probabilities.
- It is demonstrated that MTD rank histogram can diagnose over- or under-dispersion as well as miscalibration with respect to autocorrelation of scenario time series.
- Wind power scenarios generated by two very different approaches are evaluated by Energy scores, rank histograms and Brier scores.

In chapter 3:

- Wind scenarios are assessed by both reliability metrics and unit commitment simulation.
- The MTD rank histogram screens out scenario sets that would result in high costs if they were used for solving the stochastic program and the resulting solution were implemented.
- Brier scores for critical events provide ranking of remaining sets of scenarios.
- The numerical study indicates that the reliability metrics allow choice of a suitable scenario generation method without extensive simulation.

In chapter 4:

- Methods for scenario quality assessment in the context of a target optimization problem without or before explicitly solving the SP are developed, that are general enough to apply to any repetitive two-stage SP problem for which historical observations of the uncertain parameters are available.
- These methods take into consideration the impact of each scenario on the solution to the SP problem.
- EV-based scenario assessment allows comparison of the variants of a scenario generation method.
- PI-based scenario assessment is developed to compare distinct scenario generation methods.
- Random scenario based assessment is developed to evaluate the results of PI-based scenario assessment.

The limitations of the proposed approaches presented in this dissertation should be considered before application. The MTD rank histogram can result in a deceptively flat rank histogram for some combinations of parameters when assessing the scenarios directly. The emphasis of Energy Score (ES) on sharpness could introduce risk because a low ES can be obtained by sharp scenarios. The Brier score, used to evaluate wind power scenarios according to the pre-defined critical events, gives very low scores when the scenario sets are too sharp. The results of PI-based scenario assessment are hard to interpret for SP problems that have sensitive optimal objective values. The random scenario based scenario assessment could help to interpret the results, but we need to further investigate when the difference between similarity of PI-based and random scenario based approaches of different scenario sets are significant. For future research, we can further investigate how to quantitavely evaluate the PI-based scenario assessment for SP with sensitive optimal objective values. Moreover, we can test EV-based and PI-based scenario assessments with simulation studies when simulated observation values are generated with more complex correlation models rather than AR(1) or Bernoulli distributions (in the case of binary random variables). With such a test, we can evaluate the sensitivity of these approaches.

This work can be extended to multi-stage SP problems. However, there is an important qualitative difference between two-stage and multi-stage SP problems. In multi-stage SP models, scenarios have a tree structure. In two-stage SP models the only real decision is taken in the root of the tree while second-stage decisions are obtained by solving the recourse problem. In multi-stage models there is a recourse problem for every non-leaf node in the tree. Scenario tree assessment for multi-stage SP models would require the use of metrics that account for the tree structure.