Norwegian University
of Life Sciences

Master Thesis 2020 30 ECTS
Faculty of Science and Technology

Verification of Probabilistic Predictions for Reliability Analysis in the Norwegian Transmission System

Yasmin Bashir Sheikh-Mohamed
Environmental Physics and Renewable Energy

Acknowledgements

This thesis marks the end of my time as an Environmental Physics and Renewable Energy student at the Norwegian University of Life Sciences (NMBU). Through my years at NMBU I was able to explore my interests in environment, technical aid and much more. I am grateful for good friends and classmates who motivated me through this period.

First, I would like to thank my supervisors, Heidi Samuelsen Nygård and Katrine Bruvik for their immense support. Heidi gave me guidance and was always available if I needed advice. Your insightful feedback helped me elevate my work. Katrine provided me with expertise to understand the concepts and tools required to conduct the research. You went to great lengths to provide me with whatever I needed for my research. Such commitment is hard to find in supervisors. Thank you both.

I would also like to thank Statnett SF for providing me with data and access to systems. I would specifically like to acknowledge the MONSTER-team who were available if I had questions and always rushed to solve when I had technical problems.

In addition, I would also like to acknowledge my co-supervisors Oliver Tomic and Kristian Liland for being available when I had questions and also contributing my keen interest in Data Science.

Lastly, I want to thank my family and friends for support through this process.

The thesis has been carried out in collaboration with Statnett SF. Statnett has contributed information and guidance, but does not assume responsibility for this work that has been done or the results and conclusions that occur.

Abstract

As a part of the fossil fuel phase-out, Norway is investing in electrification and therefore expansion of the transmission grid. This requires efficient and accurate methods to assess long-term reliability and socioeconomic benefits of alternative expansions. The Norwegian Transmission System Operator, Statnett SF, has developed a probabilistic simulation tool, MONSTER, for reliability analysis and transmission system planning.

This thesis evaluates methods used for verification of probabilistic results and assesses their suitability for MONSTER predictions. Through a case study, the accuracy of the tool is assessed for different time spans to investigate the tools performance for short-term reliability, and therefore possibility for future application in other areas. A sensitivity analysis is also performed to assess the simulation tool's sensitivity to inputs.

After an assessment of the results from MONSTER and methods used for verification of probabilistic forecasts, Continuous Rank Probability Score (CRPS) was chosen as the main method to evaluate the accuracy of the results. Reliability diagrams and percentile diagram are used as complementary visual tools of assessment. The CRPS score from probabilistic results – in form of a probability density function – can be directly compared to the Mean Absolute Error (MAE) of point-predictions. Therefore, the point-predictions from the simulation tool are assessed using MAE.

The case study is based on the Greater Oslo Region over 9 years. The point predictions for yearly intervals have higher accuracy than for 6-month and monthly intervals. This indicates that the tool performs better for predictions with longer time spans. The resulting CRPS score indicates better accuracy for monthly predictions compared to yearly and 6-month predictions. Examining the results closer with the visual assessment tools shows that the CRPS score does not capture deficiencies in the probability distribution and has therefore computed better results for monthly predictions than expected. Use of score methods that detect probability distribution deficiencies is suggested for future evaluations. It is further concluded that the end results of this study are most sensitive to the remedial measures-input. Therefore an expanded use of this feature could result in better predictions.

Sammendrag

Som en del av utfasingen av fossile brensler satser Norge på elektrifisering. Dette krever utbygging av overføringsnettet. Nøyaktige pålitelighetsanalyser er nødvendig for å sikre pålitelig og samfunnsøkonomisk utbygging. Statnett SF har utviklet et probabilistisk simuleringsverktøy, MONSTER, for å vurdere langsiktig leveringspålitelighet ved å predikere sannsynlighet for ikke-levert energi (ILE) og kostnadene av dette (KILE).

Denne oppgaven evaluerer ILE-resultater fra MONSTER ved å først vurdere ulike metoder som brukes i dag for verifisering av probabilistiske modeller. I en casestudie blir nøyaktigheten til verktøyet vurdert for forskjellige tidsperioder for å undersøke mulighetene til å utnytte verktøyet for kortsiktige analyser. Det utføres også en sensitivitetsanalyse for å vurdere sensitiviteten til simuleringsverktøyet for endring i input-variabler.

Etter et litteraturstudie blir Continuous Rank Probability Score (CRPS) valgt som hovedmetode for å verifisere predikert Ikke-levert Energi fra MONSTER. I tillegg brukes pålitelighetsdiagrammer og persentildiagrammer, som visuelle verktøy for en grundigere analyse av resultatene. Siden CRPS kan sammenlignes direkte med gjennomsnittlig absolutt feil (MAE) for punkt-prediksjoner, blir MAE for forventet ILE fra MONSTER sammenlignet med CRPS-score av sannsynlighetskurven for ILE.

Området som analyseres i studien er Stor Oslo i en 9 årsperiode fra 2010 til 2018. Forventet ILE for årlige prediksjoner viser bedre nøyaktighet enn 6-måneders og månedlige prediksjoner. Dette indikerer at verktøyet gir mer nøyaktige prediksjoner for større tidsintervaller. CRPS-scoren viser bedre nøyaktighet for månedlige intervaller enn års-intervaller og 6-måneders intervaller. En nærmere analyse av resultatene – ved bruk av valgte visuelle verktøy – viser at CRPS ikke fanger opp mangler i sannsynlighetskurven, for eksempel ekstremutfall. CRPS har derfor beregnet bedre nøyaktighet for månedlige prediksjoner enn forventet. Bruk av score-metoder som fanger opp ekstremutfall kan bidra til bedre evalueringer av resultater for fremtidige analyser. Det konkluderes videre med at tiltak-inputen påvirker sluttresultatet mest i sensitivitetsanalysen. Derfor kan videreutvikling av denne funksjonen resultere i bedre prediksjoner.

Acronyms

CDF Cumulative Density Function

CENS Cost of Energy Not Served

CRPS Continuous Rank Probability Score

DSO Distribution System Operator

ENS Energy Not Served

GUI Graphical User Interface

IG Ignorance Score

MAE Mean Absolute Error

NCRPS Normalised Continuous Rank Probability Score

NMAE Normalised Mean Absolute Error

PDF Probability Density Function

PIT Probability Integral Transform

QS Quantile Score

TSO Transmission System Operator

Contents

1 Introduction

1.1 Motivation

The growth in unpredictable energy production and liberalisation of the electricity market create new challenges for traditional methods of planning and operation of transmission systems. Power systems are becoming complex and unpredictable due to renewable and dispersed energy entering the market (1). Alternative methods for reliability and risk analysis for transmission system planning are becoming increasingly important. Until now mainly deterministic methods have been used. In recent years, however, there has been a growing interest in probabilistic methods (2).

In 2019 ACER, EU's Agency for the Cooperation of Energy Regulators, adopted a decision for Transmission System Operators (TSOs) to develop a methodology for probabilistic risk assessment by 2027 (3). According to ENTSO-e, this is the beginning of many steps moving away from deterministic criteria for reliability (4). Probabilistic tools are already implemented in parts of North America (5). There have been some initiatives in Europe. GARPUR – a project initiated by SINTEF Energy Research with collaboration of 7 TSOs – researched alternative, probabilistic reliability criteria (6).

With a goal of increased electrification – to reduce carbon emissions – Norway is investing in expansion of the transmission grid (7)(8). According to a report by DNV GL, the electric power consumption in Norway can increase with up to 30-35 TWh in 2040 (9) from the 2017 consumption of 132.9 TWh (10). The need for a socioeconomic planning of expansion is one of the main drivers in the Norwegian TSO, Statnett's, development of a probabilistic tool for long-term reliability analysis, MONSTER (11).

So far the accuracy of MONSTER has not been systemically assessed. There is a need for general methods to evaluate results from the simulation tool. So far, most of research done on probabilistic verification methods is within weather forecasting. An assessment of probabilistic verification methods is required to develop a foundation for general evaluation methodology. Therefore, this study aims to assess probabilistic verification methods and evaluate their suitability to verify results from the simulation tool.

1.2 Problem Statements

The main aim of the research for this thesis is to assess methods to evaluate the accuracy of MONSTER's probabilistic predictions of Energy Not Served (ENS). Chosen methods will then be used to assess the tool's accuracy through a case study in the Greater Oslo Region through a period of 9 years.

The objectives of this thesis are:

1. To assess verification methods for the probabilistic results of ENS from MONSTER.
2. To examine the accuracy of the tool for different time intervals.
3. To perform a sensitivity analysis in order to examine the effect of different inputs on the predicted ENS.

The main area of application for this simulation tool is long-term reliability, typically 10 to 40 years. Even so, testing for smaller time intervals will display the tool's performance for smaller time periods and therefore whether it can be used for short term planning and operation in the future. Also, testing the tool's sensitivity to various inputs gives valuable insights on different inputs' effect on the predicted ENS.

1.3 Thesis Structure

This study is divided into a literature review of existing verification methods and a case study of the Greater Oslo Region. Chapter 2 presents theory on the Norwegian transmission system and existing reliability evaluation methods, as well as general introduction to probabilistic models and a description of MONSTER. Chapter 3 is a literature review of existing verification methods in the weather forecasting community and choice of methods for the case study. Chapter 4 describes the case study and the methodology of the research. Results and discussion are merged in chapter 5 and chapter 6 consists of conclusions and suggestions to further research.

2 Theory

2.1 The Norwegian Transmission System

The Norwegian transmission system consists of three parts; the transmission grid, the regional grid and the distribution grid. The transmission grid connects large producers and consumers across the country. It operates at high voltages – mainly between 300-420 kV, but also 132 kV – and connects to neighbouring countries' grids (12). The Norwegian TSO, Statnett, operates the transmission grid which consists of 11 000 km of high voltage lines, submarine power cables, and approximately 170 electrical substations across the country. Consumers are usually directly connected to the regional and distribution grids with lower voltage levels. Regional and distribution grids are operated by 130 Distribution System Operators (DSO) per August 2018 (13). Figure 2.1 illustrates a simplification of the power system, from production to consumers. Smaller power production plants can also be connected to the regional grid and distribution grids and larger consumers can be directly connected to the regional and transmission grids (14).

In addition to ownership over the transmission grid, Statnett is also responsible for operating the grid and ensuring quality for the electricity delivered to the consumers. The TSO is also responsible for planning and developing the transmission grid (15).

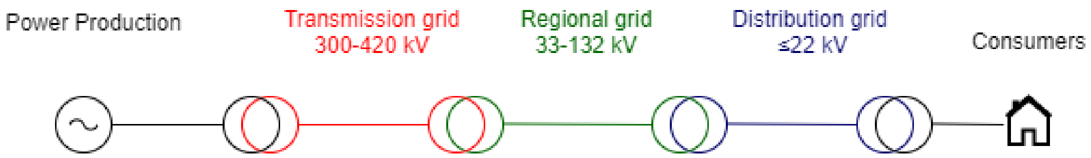


Figure 2.1: *This figure roughly illustrates the Norwegian transmission system. Inspired by (13)*

2.1.1 Transmission System Operation

The Norwegian TSO is responsible for operating the transmission grid, delivering electricity at an adequate quality, and minimising interruptions (15). The voltage level can have a deviation of maximum ± 10%, frequency has to be in the 50 Hz ± 2% range, and the voltage symmetry between the phases can have a maximum deviation of 2% (16). The TSO is also responsible for keeping the balance between production and consumption at all times (15). Since the Norwegian grid has connections to other European countries, the balance expression can be extended as shown in Equation 2.1.

$$import + production = export + consumption + losses. \tag{2.1}$$

Ensuring that Equation 2.1 is maintained at all times equates to momentary balance. In the government's regulations – relating to the system responsibility in the power system – the system operator is responsible for maintaining momentary balance (17).

2.1.2 Reliability Evaluation

Operating and planning the transmission grid includes socioeconomic beneficial planning. The TSO is also responsible for ensuring secure and reliable power systems, which includes investing in reliability. Considering that increased reliability naturally equate to higher costs, these two objectives contradict. Therefore, a TSO should always have an objective of optimising the trade-off between investment costs and reliability (2). Figure 2.2 demonstrates this trade-off.

Reliability refers to a power system's ability to deliver electricity within the given constraints. Energy is categorised into *served* and *not served*, which also incorporates electricity with lower quality than admissible levels as non-delivered energy (18). The term Energy Not Served (ENS) will be used to describe this energy.

The reliability worth is an important aspect in transmission system planning and operation. There are many existing models used to quantify reliability worth. In MONSTER the reliability worth is estimated using a cost function accordant with the Norwegian Energy Regulatory Authority (11)(19). The estimated cost of ENS will be referred to as Cost of Energy Not Served (CENS) in this paper.

The quantification of reliability worth can be used to evaluate and prioritise investments. As seen in Figure 2.2, the optimum level is the minimum total societal cost. The difficulties of estimating consumer interruption costs complicates the accurate prediction of socioeconomic costs. This affects the estimated optimal reliability level (18).

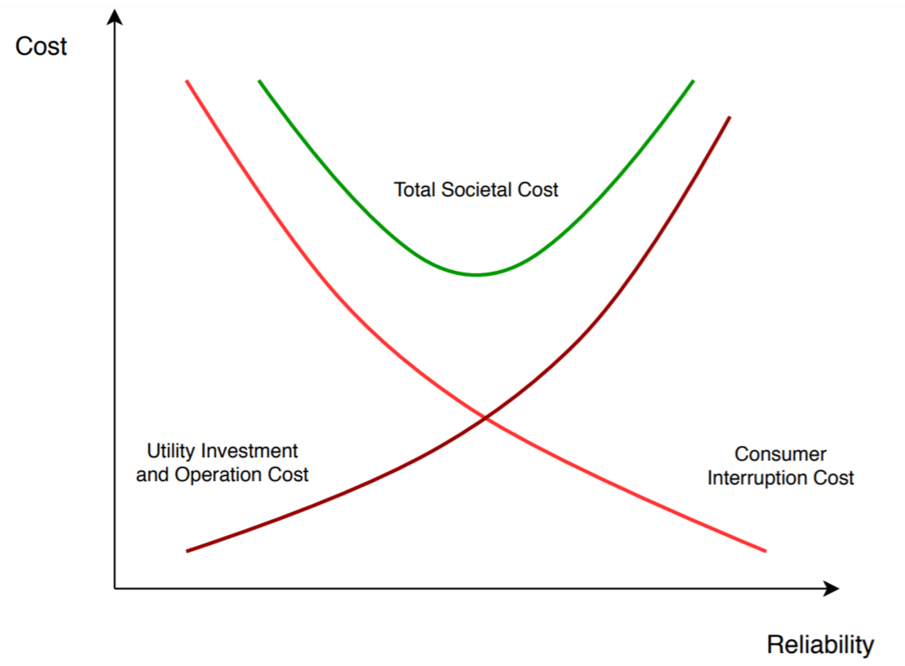


Figure 2.2: *The cost vs. reliability for investment and operation costs, and consumer interruption costs. Inspired by (18)*

2.1.2.1 N-1 Criterion

The N-1 criterion is a deterministic criterion that has provided a guideline for transmission system planning and operation to ensure reliability. The criterion states that a power system should be able to withstand disturbance of one element, for example line, transformer or generator, without transgressing the operating constraints (2).

Multiple variations of this criterion is used by European TSOs in different situations. So far, this criterion has worked to ensure reliability in transmission systems, because power systems have largely had centralised and predictable production. Deterministic criteria, like N-1, may not ensure reliability as systems grow more complex and electricity production becomes less predictable (20).

2.2 Probabilistic Models

Probabilistic modelling is a common term for models that take include the uncertainty in input variables. For complex systems including uncertainties in all parts of the model will provide a wider range of possible outcomes (21). Unlike deterministic models – that produce point-result – probabilistic models mainly produce results in the form of a probability density function (PDF). This provides more information for risk assessment and decision making (22).

The deterministic methods used for reliability analysis do not take the uncertainties of inputs into account. This disregards many possible scenarios. With the increase of variable and decentralised power production, uncertainties also increase (20). The fossil fuel phase-out, also requires increased electrification, which amplifies the need for further development of the transmission grid in Norway (23). A disadvantage with deterministic reliability analysis is that probabilities of the scenarios are not taken into account. This practice may lead to investing in reliability for unlikely events while overlooking more probable occurrences (2). As the interest in optimising the trade-off between reliable power systems and low costs investments has increased, so has the interest of probabilistic models for reliability evaluation (20).

The N-1 criterion overlooks the probability of a component failure as well as the likelihood of multiple components failing. In situations where the likelihood for failure is very low, this criterion may lead to an over-investment, while areas with higher probability of multi-component failure might be ignored. Furthermore, outages on a larger scale are usually due to multiple component failures, which proves that N-1 alone may not be appropriate to ensure system reliability. Using more secure deterministic criteria, such as N-2 or N-3, require higher investments. Utilising probabilistic methods to complement the N-1 criterion during decision making can increase system reliability and reduce costs (2).

2.3 MONSTER Simulation Tool

This subsection gives a description of the probabilistic simulation tool, that is used for this study. MONSTER is a tool developed by Statnett for probabilistic reliability analysis and is based on individual, weather dependent failure rates, with Bayesian Updating and Monte Carlo simulations (24). An overview of the simulation tool's modules is illustrated in Figure 2.3. Before describing the different modules of the tool, a description of Monte Carlo and Bayesian Methods will be given.

2.3.1 Monte Carlo Simulations

Monte Carlo Methods is generally used to describe methods that use randomness, often to solve problems in complex systems. The methods include uncertainties in inputs of the model by drawing random values for the inputs. Usually, a large number of simulations are computed. An advantage of Monte Carlo Methods is the ability to perform risk analysis in a system (25).

Given a simple model, $M(x)$ with one input, x – where x has an uncertainty represented by the probability distribution, $f(x)$ – the probability distribution $f(x)$ gives the probabilities for each possible outcome of x. In a Monte Carlo method – with n number of simulations – n random values of x are chosen based on the probability distribution $f(x)$. For each randomly chosen value, $M(x)$ is computed. To make sure the number of simulations – in other words, samples of x – are enough to be representative for the whole distribution, convergence is controlled for. Equation 2.2 illustrates the mean of simulated results, called the expected value (26). How to determine which simulation number to use depends on the convergence of the Monte Carlo simulations (27).

$$\hat{\mu}_n = E[M(x)] = \frac{1}{n}\sum_{i=1}^{n} M(x_i). \tag{2.2}$$

2.3.2 Bayesian Updating

Bayesian Updating uses Bayes Theorem, where the prior probability of an event, $P(A)$, is updated to estimate conditional probabilities based on given information, B. Bayes Theorem is stated in Equation 2.3, where $P(A|B)$ is the updated, posterior probability (28).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.3}$$

In the simulation tool, Bayesian updating is used to compute individual failure rates for components. The prior probability can either be based on general failure rates of components with similar properties or expert evaluation. Components of the same type can have different sensitivity when exposed to the same weather. To take this into account, each individual component's prior failure rate is updated by using its individual failure rate. The method is explained in more detail in (29).

2.3.3 The Simulation Tool

The simulation tool computes weather dependent failure probabilities using historical failure rates and weather data. This way failure rates for lines are increased simultaneously depending on weather conditions. This increased probability of failure is more realistic, since weather conditions affect lines' probability of failure, simultaneously (24). The failure probabilities for cables, station components, and transformers in his tool, are not weather dependent (11). This is due to low or no sensitivity to weather changes in these components (30).

Figure 2.3 depicts the tool's modules. The *Time Series Manager* computes hourly failure rates for the components in the analysis using failure statistics from FASIT – the official Norwegian database for reporting failures – and weather data. Failures for lines are divided into 8 types, as seen listed in Table 2.1. For each failure type and component, hourly time series of failure probabilities are computed in the Time Series Manager. These time series are then used in the *Monte Carlo* module, where a number of outages are drawn per simulation and distributed across the simulation period. A start time and duration of failure are also drawn (24).

The *Outage Manager* module goes through each hour in the simulation period and collects contingencies. Contingencies here are defined as instances where at least one component is unavailable. Components may be unavailable because of failures drawn in the Monte Carlo-simulations or due to maintenance, added through the Service Plan (11).

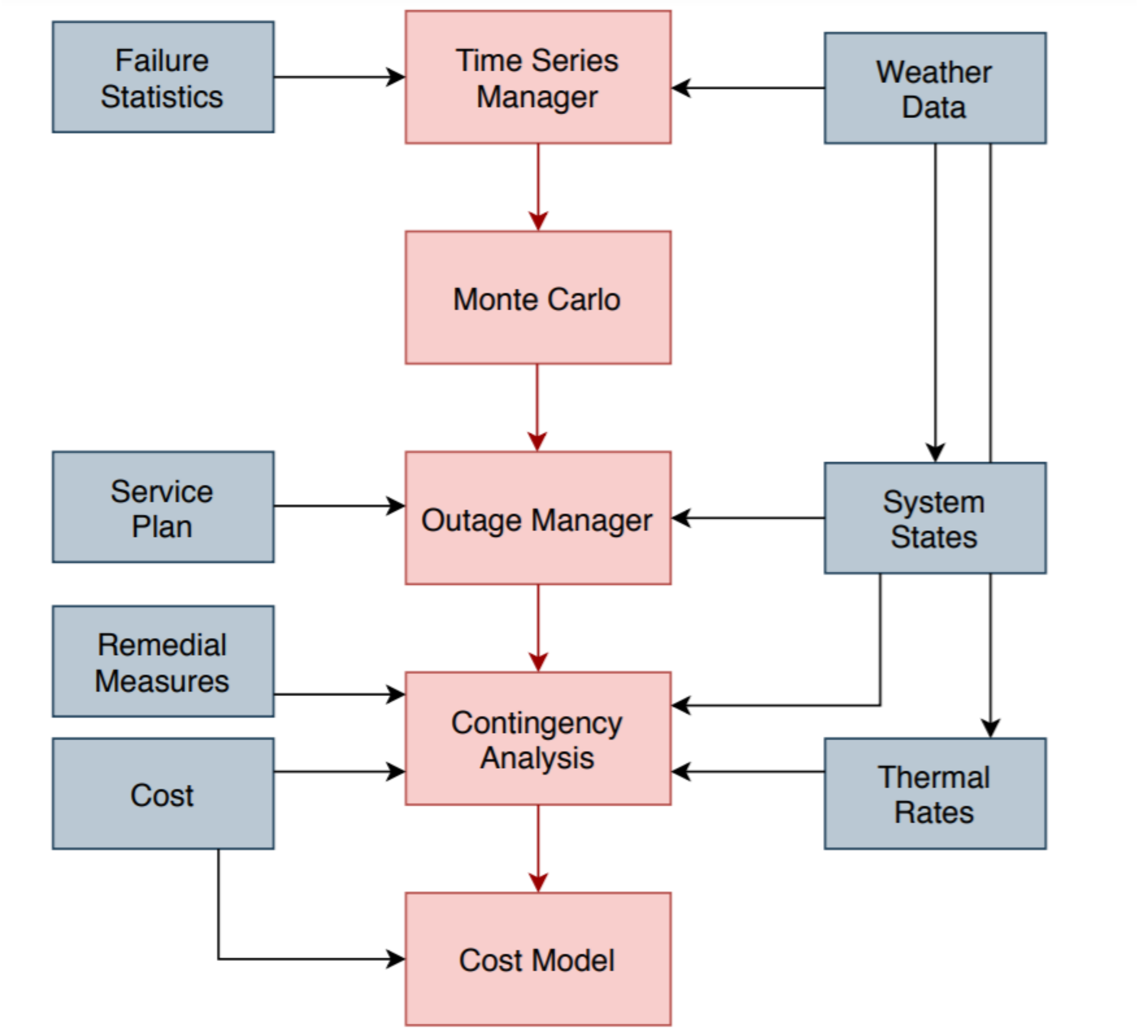
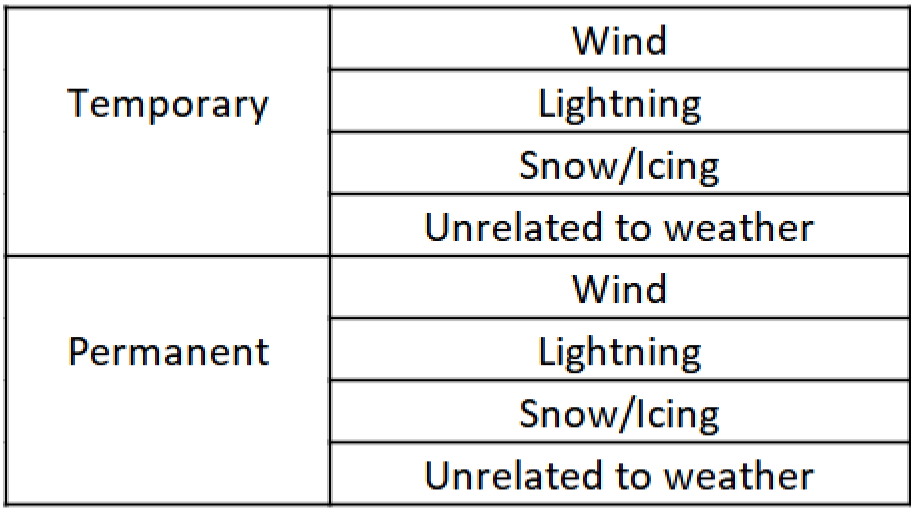


Figure 2.3: *An overview of the structure of MONSTER. Inspired by (11)*

Table 2.1: *An overview of the 8 types failure rates in lines are categorised into.*



Temporary	Wind
	Lightning
	Snow/Icing
	Unrelated to weather
Permanent	Wind
	Lightning
	Snow/Icing
	Unrelated to weather

The contingencies from the *Outage Manager*-module are analysed in the *Contingency Analysis*-module. To estimate the ENS, consequences of each contingency is assessed for the modelled system state in the time period of the contingency. The system state consists of load levels throughout the year, line transfer capacities, and flow patterns. Thermal rates are also added since temperatures affect line transfer capacities. In this module, remedial measures are also added, imitating measures taken by the system operator to prevent or reduce outages. These include changing production, moving loads etc. For simulated contingencies that lead to power outages, ENS is predicted. The Cost Model-module predicts the CENS – based on the ENS, the duration of the outage, time of the year and, the type of load – using methods proposed by the Norwegian Energy Regulator Authority (11)(19).

2.3.4 Predicted ENS from MONSTER

Cost estimations are out of the scope of this thesis. Therefore, to exclude uncertainties in estimations of the reliability worth, predicted ENS from MONSTER is analysed. The ENS results are presented as a cumulative density function (CDF) based on the computed ENS per simulation.

A cumulative distribution function, $F(x)$, is described as the probability that the predicted value, η is equal to or less than x. The definition is described in Equation 2.4. The function is non-decreasing, $\Delta F(x) \geq 0$, $F(-\infty) = 0$, and $F(\infty) = 1$ (31). An example of a CDF-plot produced by MONSTER is shown in Figure 2.4.

$$F(x) = P(\eta \leq x) \tag{2.4}$$

A point prediction for ENS is also estimated, based on the mean of computed ENS per simulation per year. This value will be referred to as *the expected value.* The calculation for the expected value, E, is shown in Equation 2.5, where y represents the year, N represents the number of Monte Carlo simulations in the prediction, and $f(x)_{ij}$ is the computed ENS for the jth simulation on the ith year.

$$E(f(x)) = \frac{\sum_{j=1}^{N} \sum_{i=1}^{y} f(x)_{ij}}{yN}. \tag{2.5}$$

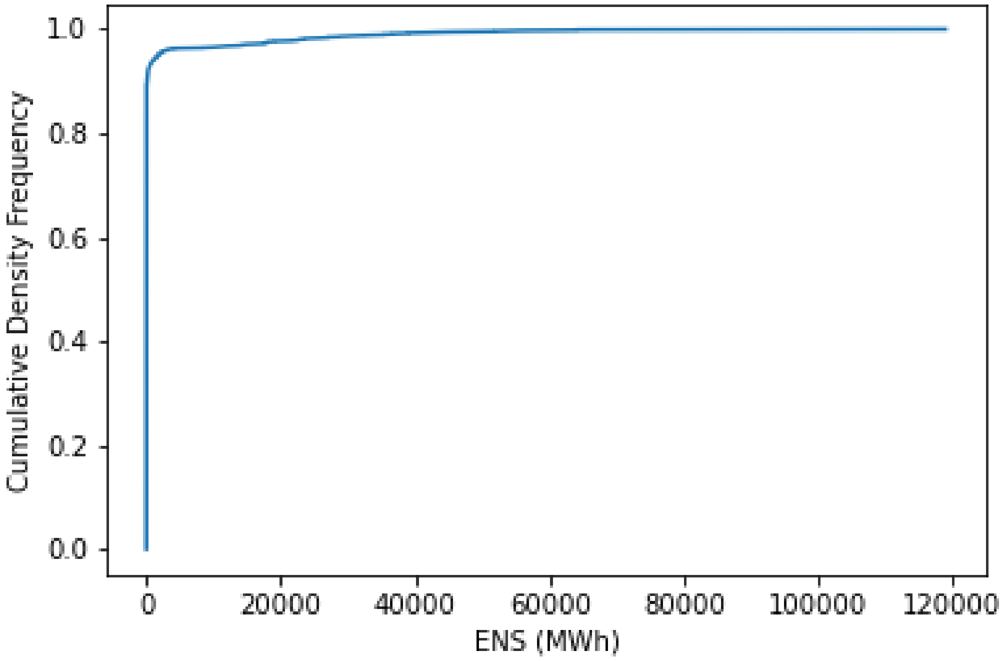


Figure 2.4: *Cumulative density plot for a MONSTER-run for year 2010. The function has a hight valye $ENS = 0MWh$, because of high probability density at this point.*

2.3.5 Extreme Values

The CDF in Figure 2.4 flattens out around $ENS = 40,000MWh$. The long, almost flat tail – caused by a very small increase in cumulative density at high ENS values – will be referred to as *extreme values*. These values are important for conducting risk analysis. However, when predicting accurate reliability worth, these extreme values will significantly affect the expected value. Considering that the expected ENS value is based on an average of all computed ENS per year per simulation for a MONSTER run, extreme values can negatively affect the expected value's accuracy.

3 Choice of Methods for Verification

Verifying probabilistic results is more complex than deterministic point-predictions. It is easier to quantify and evaluate a point-prediction against an observed value compared to a Probability Density Function (PDF) (32). This chapter will go through the theory of probabilistic verification and the process of assessing methods to evaluate the accuracy of MONSTER. First, a brief introduction will be given on probabilistic models for wind forecasting, since the wind forecasting community is leading in verification research for probabilistic results (33). Second, multiple verification methods will be discussed and evaluated to choose methods fit for evaluating ENS predictions from MONSTER.

3.1 Probabilistic models for Wind Forecasting

There are different types of probabilistic wind forecasting methods. Parametric methods are based on the assumption that the prediction frequency follows a predefined shape, for example, a Gaussian distribution (34). Non-parametric methods' PDFs do not follow a predefined shape. An example of a non-parametric method is ensemble forecasting. Ensemble forecasting is a scenario-based forecasting method that takes into account the uncertainties in the input variables. A PDF is produced based on the results from the number of ensembles used for the model (35). The process of this type of model is presented in Figure 3.1. Ensemble models are not necessarily based on probabilistic criteria, since ensemble models can consist of a set of deterministic forecasts, but they do produce a PDF-predictions (34).

3.2 Required Qualities and Evaluation Diagrams

There are multiple requirements mentioned in various literature for probabilistic forecasts (36)(33). This paper will discuss reliability, sharpness, and resolution. Reliability, also called calibration, refers to the model's ability to consistently predict results similar to historical observations (33). Resolution is a forecast's ability to predict different predictions based on different input variables (32).

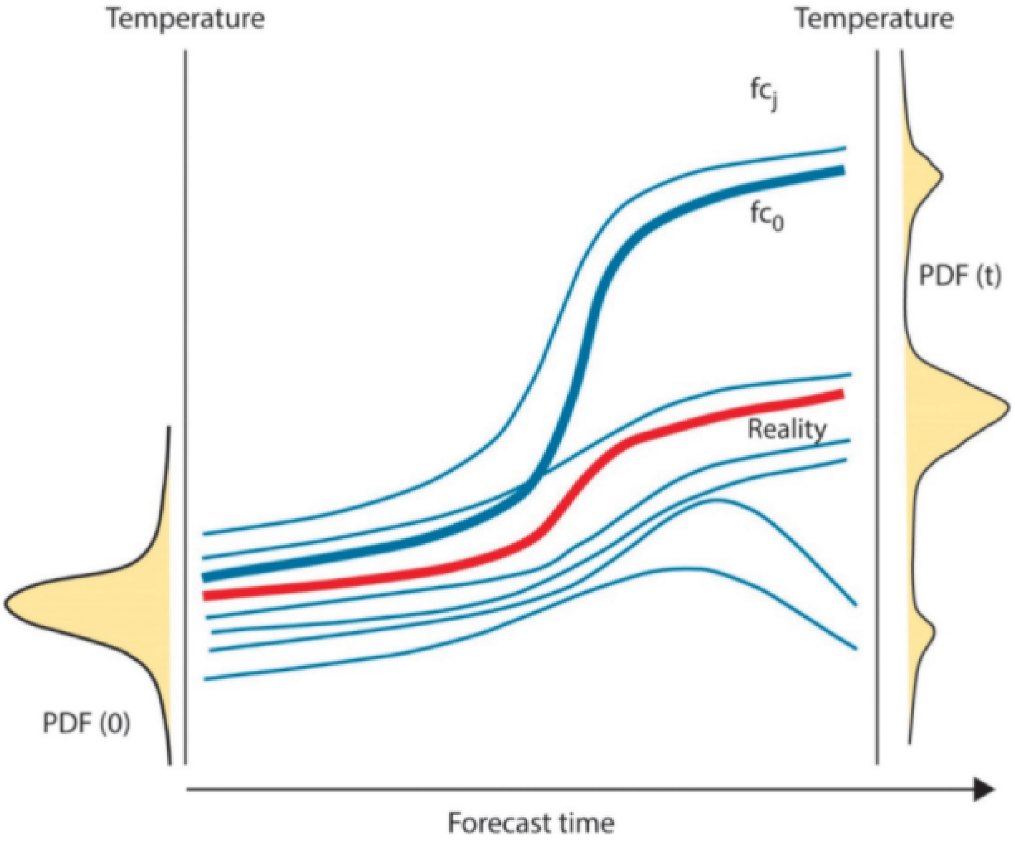


Figure 3.1: *Illustrates the process of ensemble modelling (35). Reprinted with permission* [1].

Sharpness is the attribute of concentration of the probability distribution. For example, as seen in Figure 3.2, both PDFs are centred around the same value of the parameter of interest, x. However, one PDF is more concentrated than the other and therefore gives a more informative prediction. Contrary to the reliability, the sharpness attribute does not depend on the historical observations, but only the predicted PDF's shape (33).

There are multiple existing methods for assessing these qualities. Score methods, which will be discussed more in-depth in the next subsection, give quantitative measures of the prediction quality. Visual assessment tools are also used, such as reliability and sharpness diagrams (33).

[1] This Figure was published in Sub-seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting, 1st Edition, Andrew Robertson and Frederic Vitart, Page 38, Copyright Elsevier (2018)."

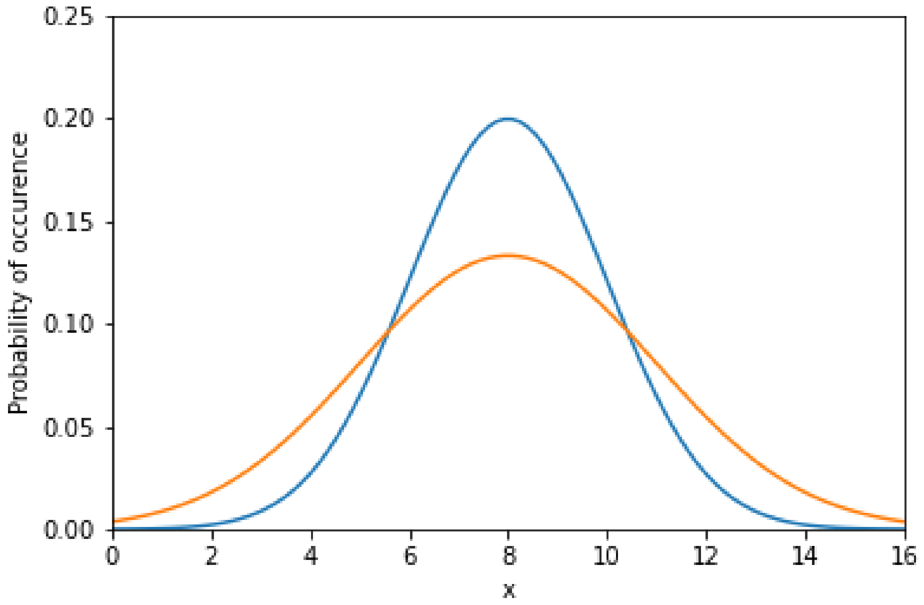


Figure 3.2: *The figure displays the PDFs for two different estimates, with the same central value but different sharpness.*

Reliability diagrams are visual assessment tools used to evaluate probabilistic predictions. Using reliability diagrams provides information on possible error causes in the prediction. In a reliability diagram the prediction percentiles, $\hat{q}^{\alpha}$, for different nominal levels, α, are plotted against observed frequency. For example, for a quantile forecast with a nominal level of $\alpha = 0.5$, 50 % of the PDF is below the $\hat{q}^{0.5}$-value. For a perfect probabilistic prediction, 50 % of the observed values, y, should also be below $\hat{q}^{0.5}$. As shown in Figure 3.3, the stapled line in the diagonal shows the ideal prediction. Evaluating a reliability diagram can give information on the effects of small sample sizes, lack of resolution and, the predictions' reliability (37).

The observed relative frequency, a^{α}, per nominal value is given by:

$$a^{\alpha} = \frac{\sum_{i=1}^{N} \xi_i^{\alpha}}{N}. \tag{3.1}$$

ξ_i^{α} is the indicator given by Equation 3.2 and N is the number of predictions.

$$\xi_i^{\alpha} = \begin{cases} 1, & y < \hat{q}^{\alpha} \\ 0, & otherwise \end{cases} \tag{3.2}$$

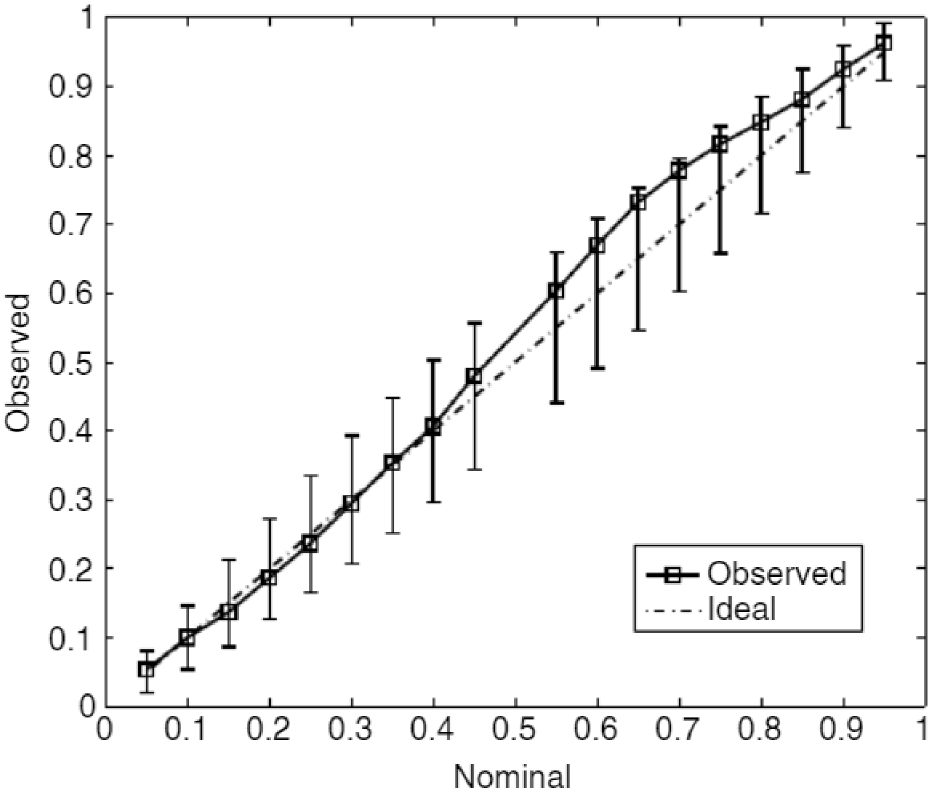


Figure 3.3: *An example of a reliability diagram. The nominal levels, in the x-axis, are plotted against the relative, observed frequency of the observed values (38). Reprinted with permission* [2].

Sharpness diagrams for forecast verification can be produced by plotting the width of interval forecasts, $\hat{I}^{\beta}$ for different nominal coverage rates. The sharpness, δ^{β} of an interval forecast $\hat{I}^{\beta}$ is given by Equation 3.3 (39).

$$\delta^{\beta} = \hat{q}^{(1-\frac{\beta}{2})} - \hat{q}^{(\frac{\beta}{2})}. \tag{3.3}$$

[2] This figure was published in Quarterly Journal of the Royal Meteorological Society, Vol 136, Pierre Pinson, Patrick McSharry and Henrik Madsen, Non-parametric probabilistic forecasts of wind power: required properties and evaluation, Page 88, Copyright John Wiley and Sons (2010).

3.3 Score Methods

Scoring methods quantify probabilistic predictions' quality by giving a numerical score (40). Score methods make an overall assessment of prediction quality (41). This simplified way of assessing a model's performance may not be informative enough (36). Some scoring methods can be decomposed to assess different attributes separately (42) (43). In this subsection, multiple score methods will be introduced and discussed.

3.3.1 Brier Score

The most widely used Brier Score evaluates probabilistic predictions for binary results. The score value can be compared to the mean squared error (36) and is one of the oldest probabilistic verification methods (43). Although mostly used for binary outcomes, the original definition allows it to be used for non-binary (44). However, only the equation for binary results is presented in this thesis.

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2. \tag{3.4}$$

Here, N is the number of instances, f_i is the probability of the event occurring for instance i, and o_i is a binary value that depends on whether the event occurred in instance i or not, as shown in Equation 3.5.

$$o_i = \begin{cases} 1, & occurrence \\ 0, & non-occurrence \end{cases} \tag{3.5}$$

3.3.2 Continuous Rank Probability Score

Continuous Rank Probability Score (CRPS) can be used for continuous and discrete results. The method compares the predicted and observed CDFs. Figures 3.4 and 3.5 illustrate the probability and cumulative probability of predicted and observed values, respectively. The vertical line in Figure 3.4 represents the observed value of the parameter of interest. For Figure 3.5, the cumulative probability $F_{obs} = 0$ up to the observed value, where $F_{obs} = 1$. CRPS penalises based on how much the predicted cumulative density function (CDF) deviates from the observed CDF (41). This allows CRPS to evaluate the prediction's reliability and sharpness. Equation 3.6 presents the CRPS calculation for one instance.

$$CRPS(F, x_{obs}) = \int_{\infty}^{-\infty} [F(x) - F_{obs}(x)]^2 \, dx. \tag{3.6}$$

$F(x)$ and $F_{obs}(x)$ are cumulative frequency distributions of the probabilistic prediction and the historical observation. For observed value, x_{obs}, $F_{obs}(x)$ is:

$$F_{obs}(x) = \begin{cases} 0, & x < x_{obs} \\ 1, & x \geq x_{obs}. \end{cases} \tag{3.7}$$

The mean CRPS of N instances is given by:

$$\overline{CRPS} = \frac{1}{N} \sum_{i=1}^{N} CRPS(F_i, x_{i,obs}). \tag{3.8}$$

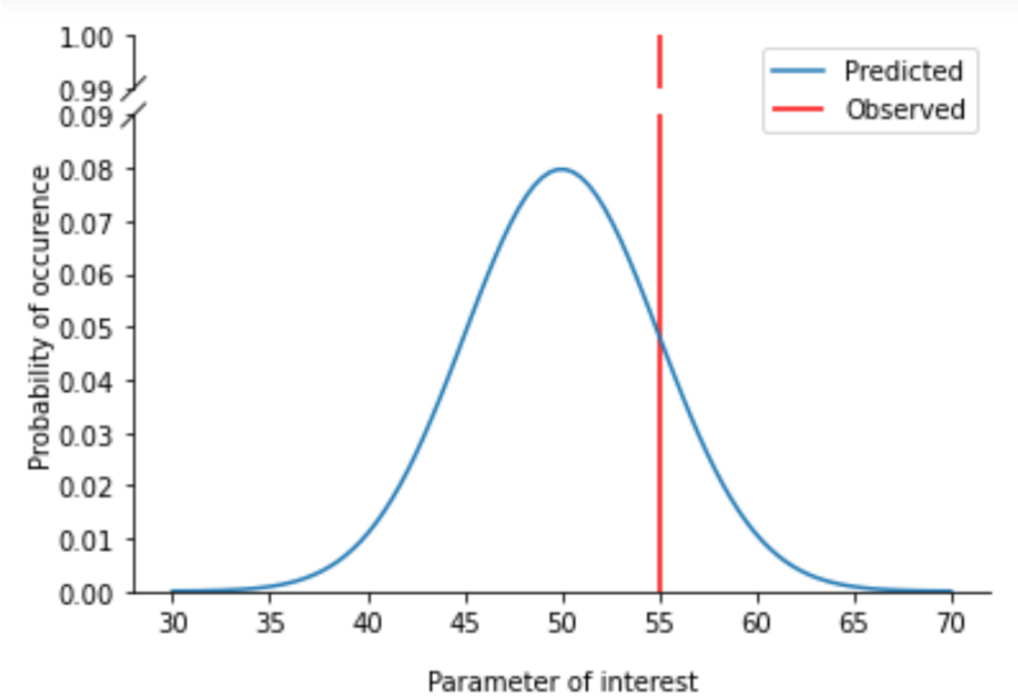


Figure 3.4: *The PDF of a prediction of parameter of interest, x. The vertical line represents the observed value at x=55.*

3.3.3 Ignorance Score

The Ignorance Score (IG) is a score method that is also used for continuous results (33). With the predicted PDF, $f(x)$, and observed value x_{obs}, the ignorance score for N observations can be calculated by:

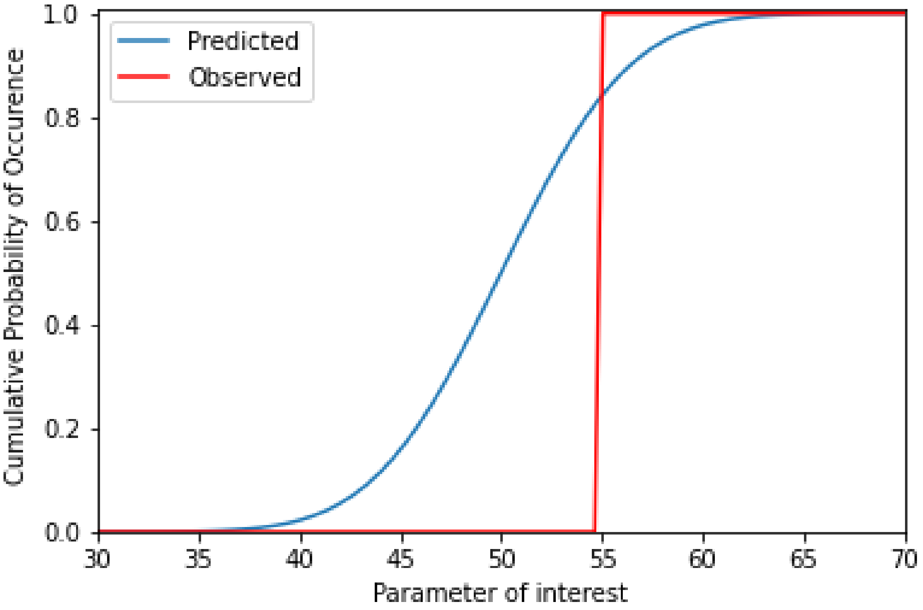


Figure 3.5: *The CDF of a predicted and observed value. The CRPS score penalises based on the deviation between the two CDFs.*

$$IG = -\frac{1}{N}\sum_{i=1}^{N} log(f_i(x_{i,obs})). \tag{3.9}$$

IG gives a higher penalty to predictions that deviate from the observed value, than CRPS. However, because of its logarithmic scale, IG cannot be implemented on PDFs with null probabilities. The score cannot be normalised either (33).

3.4 Evaluation of Methods for Case Study

Predictions from MONSTER are simulations over a longer time period and not forecasts. However, the main idea in the verification methods for probabilistic forecasts is comparing PDF-predictions to point observations. Therefore, the same methods, with some adjustments, can be used to evaluate ENS results from MONSTER.

From the score methods discussed, CRPS is the best fit to use, since it can be utilised for continuous data. Both Brier and CRPS can be decomposed into reliability, resolution and uncertainty terms (42) (33), which gives more information on prediction quality. Although the IG can be utilised for continuous results, it cannot be decomposed or normalised (33). An other possible challenge with the IG is it's inability to calculate a score for null-probabilities. If the observed value, x_{obs} has a predicted probability $f(x_{obs}) = 0$, IG will not be able to compute a score. This will become challenging for results from MONSTER, since the PDF has many points where $f(x) = 0$, as illustrated in Figure 3.6. The figure is a histogram plot where the PDF is grouped by a width of $w = 10MWh$. As seen in the histogram, there are many points where the PDF has null values.

Since CRPS can be directly compared to MAE for point predictions, using CRPS for this study enables direct comparison of the MAE from MONSTER's point predictions and the CRPS from the predicted CDF.

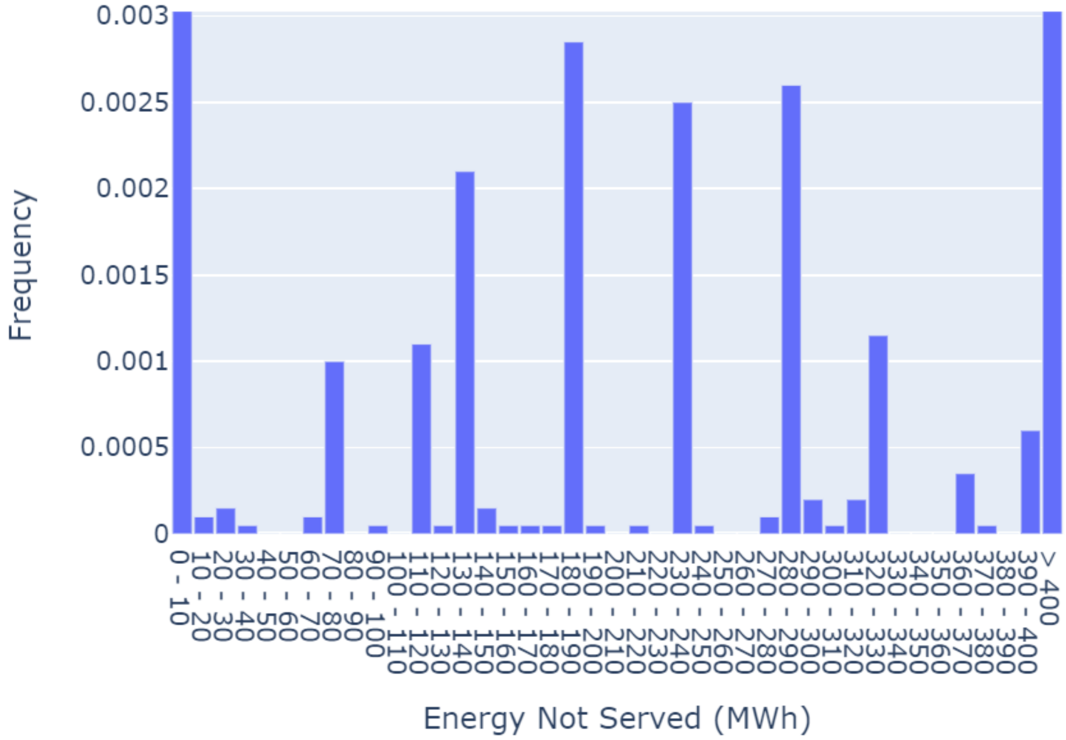


Figure 3.6: *This figure illustrates the PDF of a MONSTER-run for year 2010. The histogram is zoomed so the full length of* $0MWh \leq ENS < 10MWh$ *and* $ENS \geq 400MWh$ *are not included.*

4 Case Study

This chapter presents an outline of the area of analysis for the case study and the methodology. The methodology consists of a description of how MONSTER-runs are carried out for this study as well as data retrieval. In addition, the type of MONSTER-predictions that are made to examine the accuracy and sensitivity of the tool are described. Lastly, the methods chosen to evaluate the results of the case study are discussed.

4.1 Description of Case Study

The area of analysis is the Greater Oslo Region in the time period 2010 to 2018. Lines, cables, and components in electrical substations, as shown in Figure 4.1 are included in the study. The 7 stations that are included in the analysis are represented by red nodes in Figure 4.1. Lines and cables are represented by solid and dashed red lines, respectively. All components in the stations are included in the analysis. Transformers and station components will be collectively referred to as *station components* in this paper.

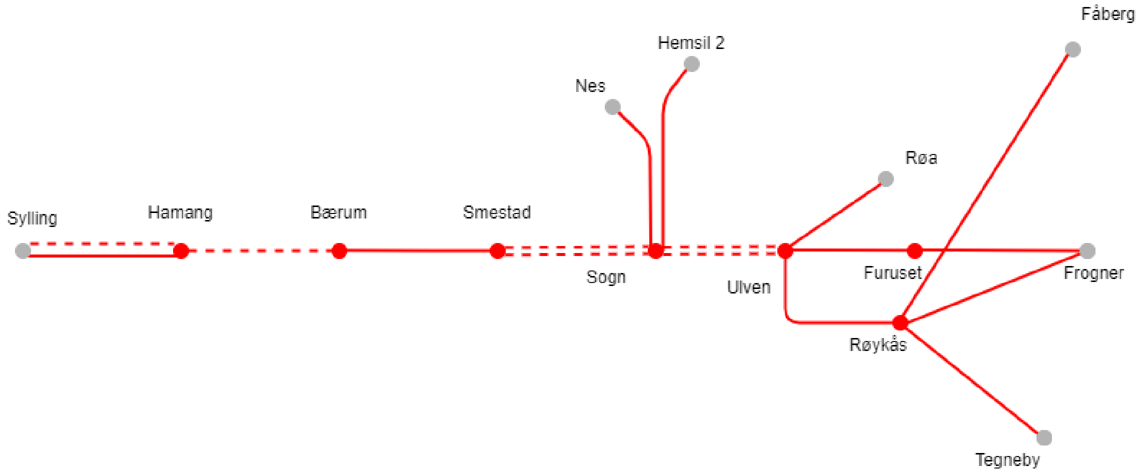


Figure 4.1: *This figure illustrates the components in this analysis. Dotted lines illustrate cables and solid lines illustrate transmission lines. The nodes represent electrical substations. The grey coloured stations are not included in the analysis.*

4.2 Methodology

4.2.1 Simplifications and Limitations in the Analysis

To test the accuracy of the tool, the input variables for this analysis have been set as similar to the historical conditions as possible. However, some simplifications were made. Since there have not been significant changes in the system state in the area and time of analysis, the modelled system state for 2019 is used for all runs. Service plans used for the analysis are the exact same as the historical service plans.

The available weather data to run MONSTER is at the moment up to 2014. Therefore, the analysis made for monthly and 6-month intervals are only for 2010 to 2014. Therefore, predictions for 2015 to 2018 uses weather data from 1980 to 2014.

4.2.2 Setting up MONSTER-runs

To set up a MONSTER-run, the network configurations are modelled in the Case Set-module. Multiple case files that represent the transmission system, and roughly the changes in power flow throughout the year. These case sets are used to map the changes in power flow throughout the year in the Case Time Mapping-module where one case file is chosen to represent the power flow per date. Electrical power usage is usually higher during the winter and lower during the summer. The power flow – modelled in the Case Time Mapping – affects the predicted consequence of a simulated contingency. For example a simulated contingency during winter will have a higher ENS compared to the same contingency with the same duration during summer. The same case time mapping is used for all MONSTER runs in the analysis, because yearly variations in power flow has been relatively similar Greater Oslo during the analysis period. Both the Case Set and Case Time Mapping are set up by Statnett, since this is outside of the scope for this thesis.

The Probability Set-module is where the probabilities of failure per component per hour are computed. All components for the analysis are added as well as all possible years of weather data. The available dates are January 2nd, 1979 to February 28th, 2015. The probability of failure per hour for all components in this time period are calculated in the Probability Set.

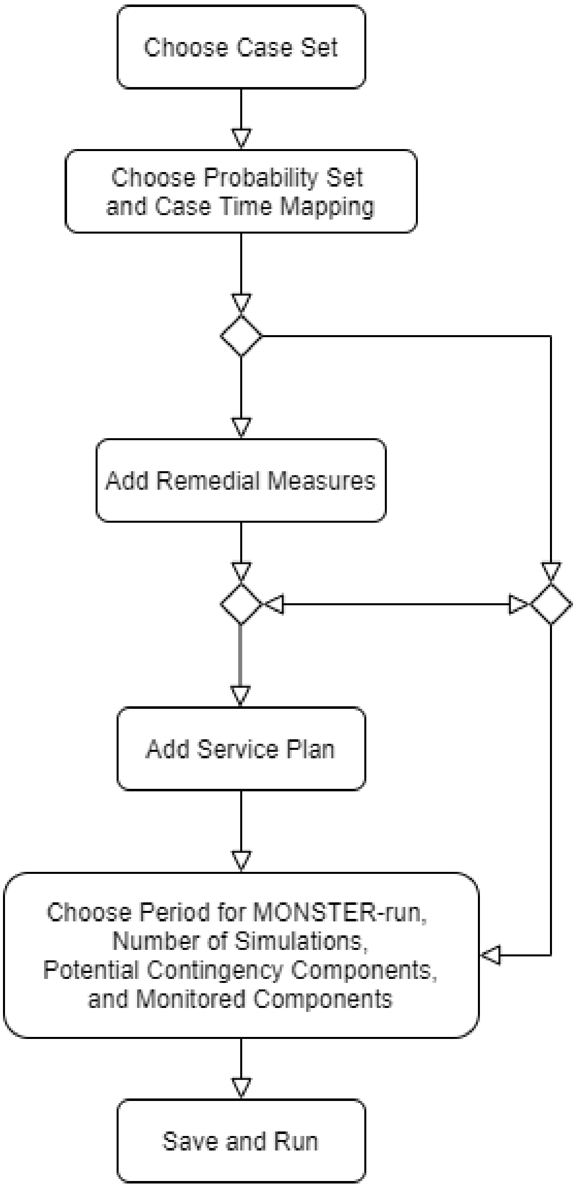
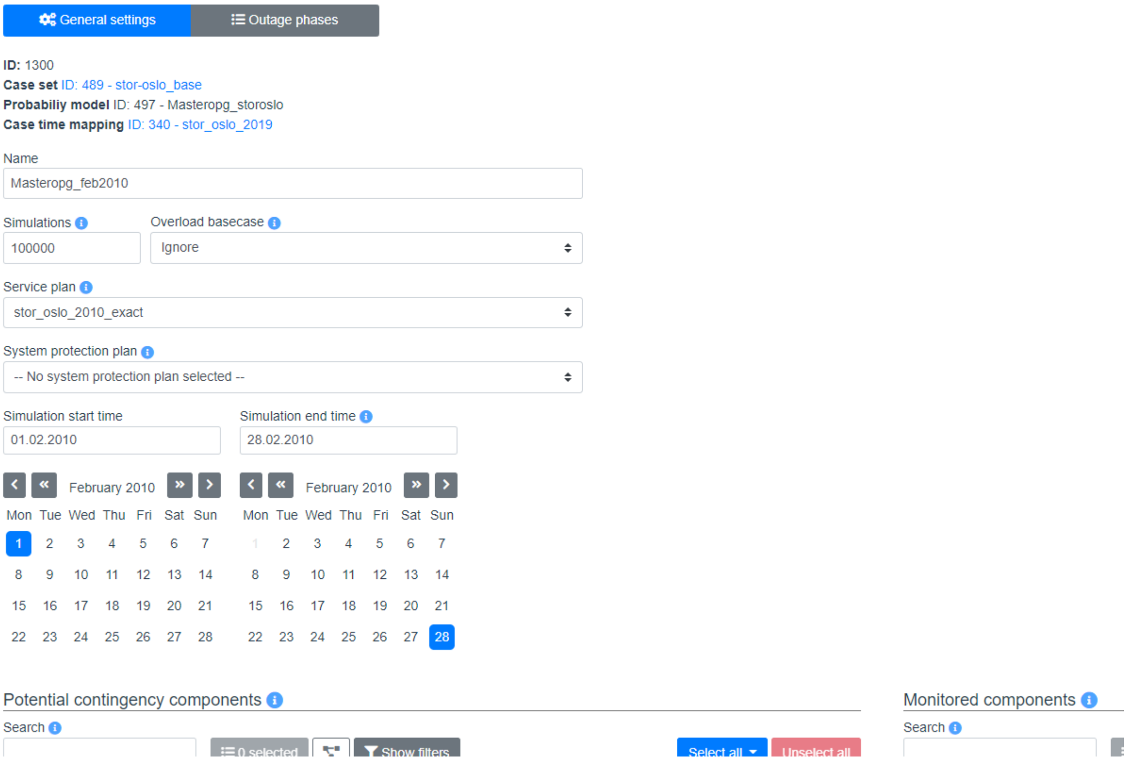

Choose Case Set
Choose Probability Set and Case Time Mapping
Add Remedial Measures
Add Service Plan
Choose Period for MONSTER-run, Number of Simulations, Potential Contingency Components, and Monitored Components
Save and Run

Figure 4.2: *Process diagram that demonstrates the process of setting up a MONSTER-run in this study. Adding service plan and remedial measures is optional.*

To get an ENS prediction a MONSTER-run is set up as illustrated in Figure 4.2. The process diagram shows how MONSTER-runs have been set up for this study, once Case Set, Case Time Mapping, Probability Set, and Service Plans were set up. In the Graphical User Interface (GUI), a Case Time Mapping and Probability Set are chosen first. All runs done for this analysis use the same case time mapping. Different Probability Sets are used for runs with changed failure rates. Figure 4.3 shows the GUI where further inputs are added. A service plan for the selected time period is added. For each MONSTER-run, the exact service plan for the selected



ID: 1300
Case set ID: 489 - stor-oslo_base
Probabiliy model ID: 497 - Masteropg_storoslo
Case time mapping ID: 340 - stor_oslo_2019

Name: Masteropg_feb2010

Simulations: 100000

Overload basecase: Ignore

Service plan: stor_oslo_2010_exact

System protection plan: -- No system protection plan selected --

Simulation start time: 01.02.2010

Simulation end time: 28.02.2010

Figure 4.3: *Snapshot of the MONSTER-run page*

period and the exact weather data is added. For example, in a MONSTER-run for February 2011, the exact service plan of 2011 and weather data for February 2011 are added. The tool will therefore use the failure probabilities computed in the Probability set for February 2011 as well as service plan for February 2011. All components for the analysis are added as *potential contingency components* and *monitored components*. Potential contingency components are the components that can get failures in the simulations. Monitoring components in the MONSTER-run allows overloads to be registered and therefore handled in the simulations. Components that are not in the Probability Set cannot be added as potential contingency components.

4.2.3 Choosing Number of Simulations

As illustrated in the snapshot of the GUI in Figure 4.3 and the process diagram in Figure 4.2, the number of simulations for the MONSTER-run is also asserted. The numbers of simulations used in this study have been chosen based on an analysis of how well the simulations converge. For each type of run, a simulation number was chosen and the predictions' convergence plots was evaluated. Figure 4.4 illustrates the convergence plot of a MONSTER-run with 20 000 simulations. If an instance had not converged, a higher simulation number was chosen for all predictions of the same type and the process was repeated. This way, convergence was controlled for all runs before data retrieval. For runs using weather data for a year or 6 months, 20 000 simulations have been used. For the yearly 2015-2018 runs, using weather data for 35 years, 5000 simulations have been chosen, and runs for monthly predictions use 100 000 simulations.

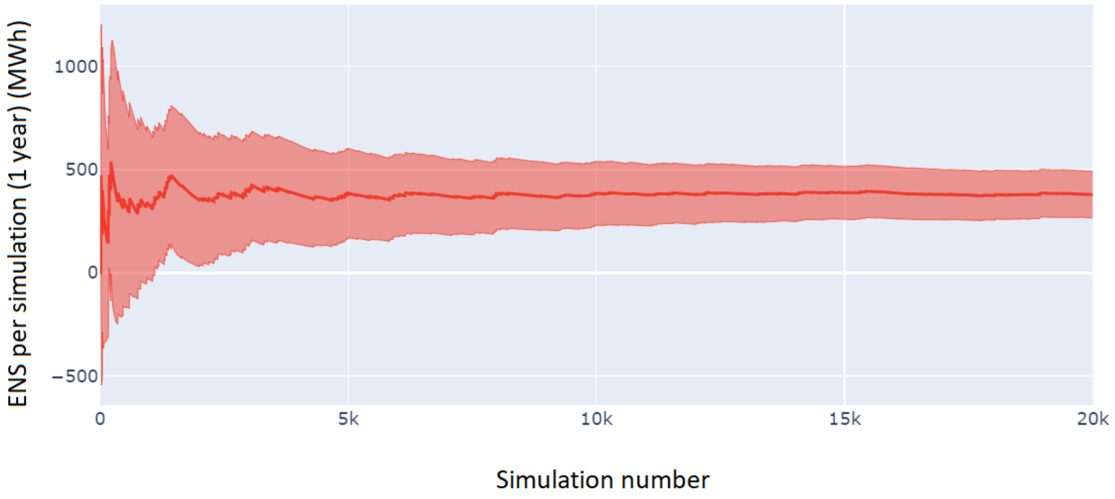


Figure 4.4: *Convergence plot for a MONSTER-run for 2010.*

4.2.4 Remedial measures in MONSTER-runs

There are multiple remedial measures that can be added in MONSTER. The remedial measure used in this study is *move load*. The System Control Centre can use the regional transmission system as an alternative transmission method, in case of a contingency. For example, in the case of a contingency in the the electrical substation, *Smestad*, leading to the station not being able to transmit power to loads supplied by this station, power can be supplied from other electrical substations and through the regional system to those loads. The amount of power that can be transmitted this way depends on the capacity of the regional transmission system and the load flow at that exact time.

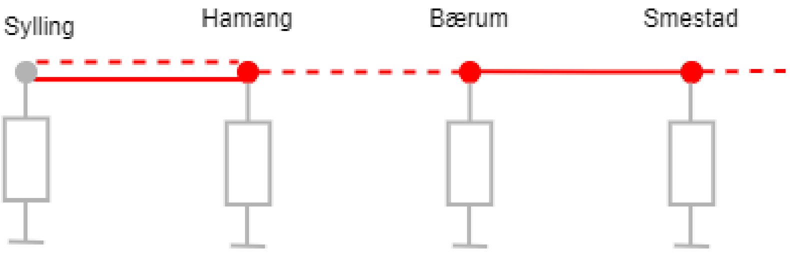
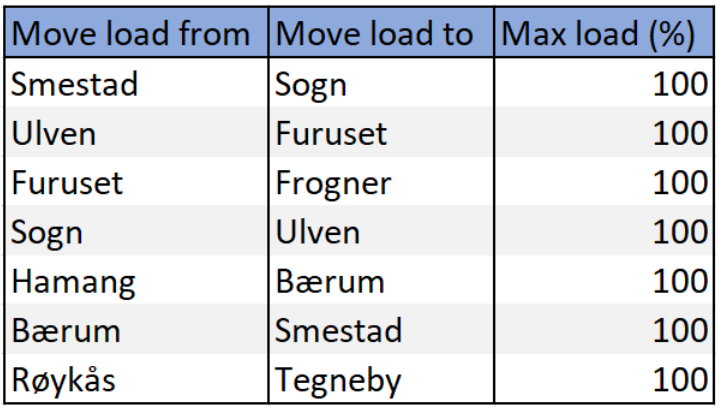


Figure 4.5: *This figure shows part of the area of analysis. Regional and distribution systems and consumers are all modelled as a load connected to the electrical substation in MONSTER.*

The regional and distribution systems are not modelled in MONSTER. Everything below the transformation station is modelled as a load as seen in Figure 4.5. To imitate the use of the regional transmission system in case of a contingency, *move load* is used as a remedial measure in MONSTER. Since the move load-measure in MONSTER cannot be customised for load flow the same maximum load can be moved regardless of the system state at the time of the contingency.

For each electrical substation in this study *move load*, has been added as a measure as listed in Table 4.1, where 100% of the load can be moved to one neighbouring station.

Table 4.1: *The table shows the specific remedial measures added in the MONSTER runs for this study. Max load is set as the maximum load that can be transferred from one station to an other in the model.*



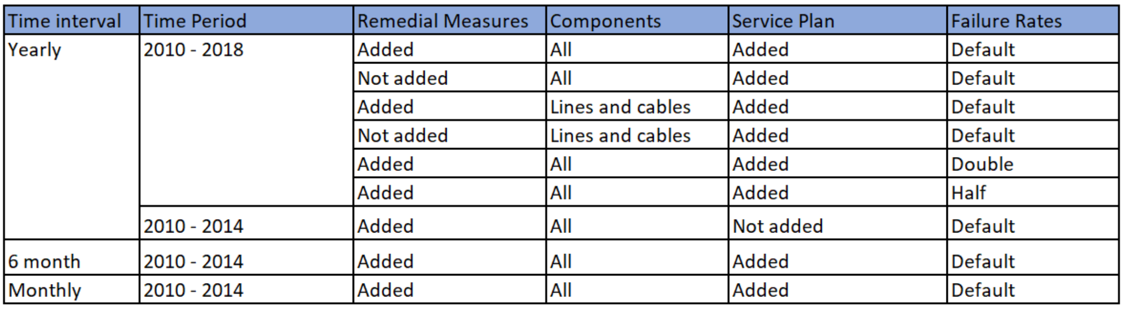
Move load from	Move load to	Max load (%)
Smestad	Sogn	100
Ulven	Furuset	100
Furuset	Frogner	100
Sogn	Ulven	100
Hamang	Bærum	100
Bærum	Smestad	100
Røykås	Tegneby	100

4.2.5 MONSTER-runs for Time Intervals and Sensitivity Analysis

To test MONSTER's accuracy for different time intervals, 3 time intervals have been chosen; year intervals, 6-month intervals and month intervals. Because of the unavailability of weather data for the years 2015 to 2018, the yearly runs made for this period use weather data from 1980 to 2014. For the 6-month intervals, the periods are divided into summer and winter, where April to September are chosen as summer months and October to March are chosen as winter months.

For the sensitivity analysis, the yearly intervals with all inputs, is used as a *base case*. Predictions are run removing different inputs, such as remedial measures and service plan to examine their effect on the predicted ENS. This analysis gives insights on the tool's sensitivity to the these inputs. MONSTER-runs are also made with changed failure rates. Since the failure rates for lines and cables are low compared to failure rates for station components, failure rates in this part of the analysis is changed only for station components. Table 4.2 gives an overview of all predictions made in this study.

Table 4.2: *This table provides an overview of all the runs made for this study.*



Time interval	Time Period	Remedial Measures	Components	Service Plan	Failure Rates
Yearly	2010 - 2018	Added	All	Added	Default
		Not added	All	Added	Default
		Added	Lines and cables	Added	Default
		Not added	Lines and cables	Added	Default
		Added	All	Added	Double
		Added	All	Added	Half
	2010 - 2014	Added	All	Not added	Default
6 month	2010 - 2014	Added	All	Added	Default
Monthly	2010 - 2014	Added	All	Added	Default

4.2.6 Data Retrieval and Processing

Historical data were retrieved from the Norwegian database for observed failures, FASIT. Failures for the selected components that have not lead to interruptions are filtered out. The data has then been grouped by date so that ENS per date is summed in case there are multiple interruptions on the same date. This results in a data frame with time series with the total daily ENS for the area of analysis. To analyse failures more closely, failures that have lead to interruptions were retrieved together with power and duration of the interruption.

The predictions from the MONSTER-runs were retrieved from the GUI, where computed ENS per year, per simulation is sorted from lowest ENS to highest. This is used to plot the CDF graph. Figure 4.6 shows a plot of the raw data as a scatter plot. Since the data frame does not have points for cumulative probability for ENS levels between the simulation points, this is added by grouping data in bins and recalculating cumulative frequencies for each bin. Figure 4.7 shows the scatter plot of the grouped data with a bin width of 1 MWh. The smaller the bin width used for the analysis, the better resolution. Larger bin widths may give a rough estimate of the numerically calculated CRPS. Therefore, the smallest possible bin width is used for this study.

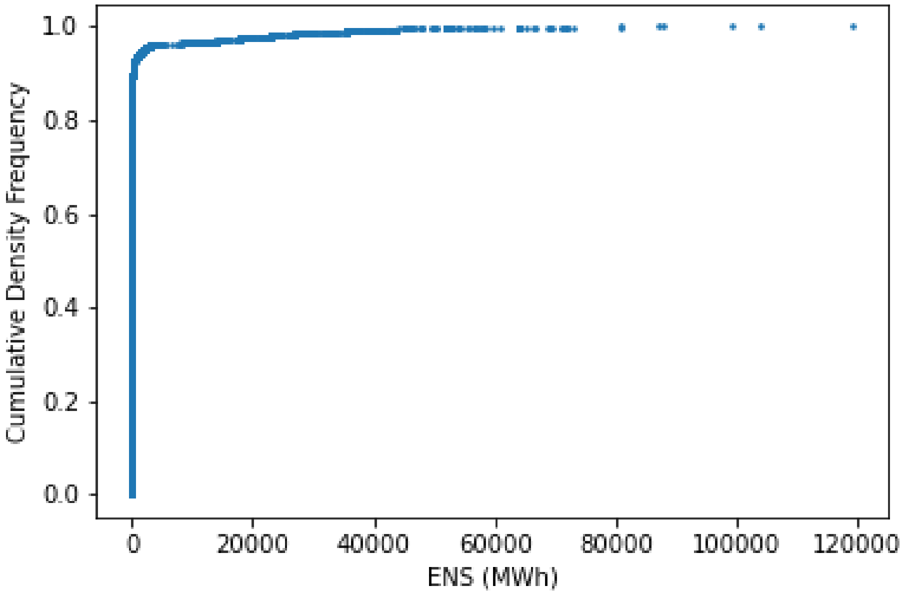


Figure 4.6: *Scatter plot of the raw data from a MONSTER prediction. There are parts of the x-axis, especially for higher values, without values*

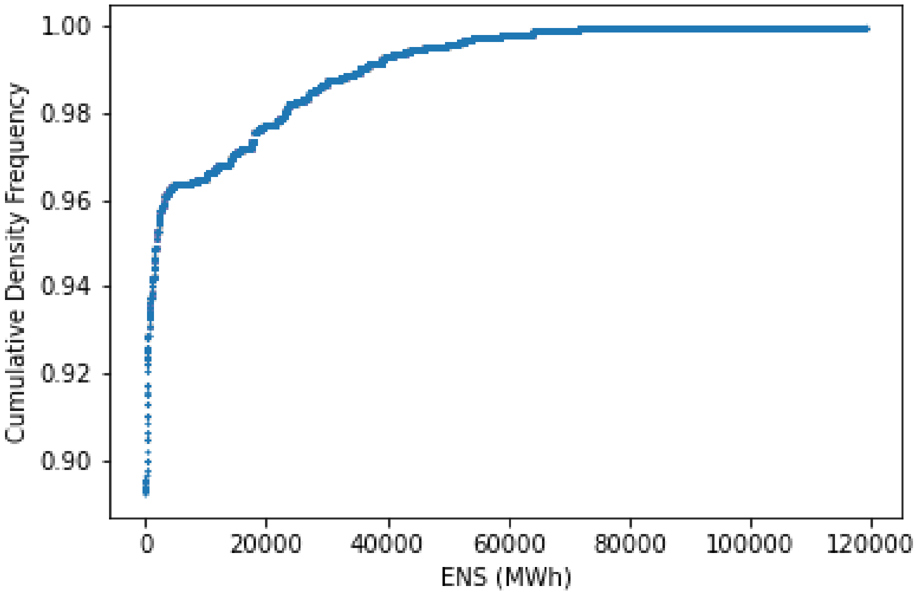


Figure 4.7: *Scatter plot of grouped data with bin width, w=1 MWh.*

4.2.7 Estimating the Accuracy of predicted ENS

When calculating CRPS numerically Equation 4.1 is used, where N is the number of instances, M_j is the number of bins in the jth instance, and x_i is the mean value of the bin interval in the jth instance. For example, bin[0,1000] MWh would have a $x_i = 500MWh$. The number of bins are different per prediction, depending on how big the interval for simulated ENS values is. However, bin width is controlled for and all calculations are done with the same bin width, $w = 1MWh$.

$$\overline{CRPS} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M_j} w[F_j(x_i) - F_{obs,j}(x_i)]^2. \tag{4.1}$$

$F_j(x_i)$ is the cumulative density for instance j at point x_i, w is the bin width and $F_{obs,j}(x_i)$ is the observed cumulative density. The normalised CRPS score, NCRPS, is given by Equation 4.2, where $ENS_{obs,j}$ is the observed ENS for instance j.

$$\overline{NCRPS} = \frac{\sum_{j=1}^{N}\sum_{i=1}^{M_j} w[F_j(x_i) - F_{obs,j}(x_i)]^2}{\sum_{j=1}^{N} ENS_{obs,j}}. \tag{4.2}$$

To compare the accuracy of the CDF predicted by MONSTER to the accuracy of the expected value, the Mean Absolute Error (MAE) of the expected value is estimated. This is given in Equation 4.3, where ENS_{pred} is the expected value of ENS predicted by MONSTER and ENS_{obs} is the observed ENS value. The normalised MAE is given in Equation 4.4.

$$MAE = \frac{1}{N}\sum_{j=1}^{N} |ENS_{pred,j} - ENS_{obs,j}|. \tag{4.3}$$

$$NMAE = \frac{\sum_{j=1}^{N} |ENS_{pred,j} - ENS_{obs,j}|}{\sum_{j=1}^{N} ENS_{obs,j}}. \tag{4.4}$$

Instead of sharpness diagrams, plots with percentile bands are used to visually assess the results from MONSTER. This is because the results generally have a high cumulative probability at $ENS = 0MWh$. Therefore, using Equation 3.3, the sharpness, δ^{β}, of an interval prediction, $\hat{I}^{\beta}$, would be equal to the percentile prediction, $\hat{q}^{(1-\frac{\beta}{2})}$, since $\hat{q}^{(\frac{\beta}{2})} = 0$ for all predicted instances in this study.

Reliability and percentile diagrams are used in this study, both for the purpose of evaluating whether they can be used to assess MONSTER-predictions and also as complementary tools to evaluate the results from the case study.

A suggested way of exclude extreme values is by using a confidence interval of the CDF instead of the full CDF. This will be tested for the yearly predictions, where the accuracy of the full density predictions will be compared to the accuracy of an interval prediction with 95 % confidence interval.

4.2.8 Percentile Diagrams

Plots with percentile bands, as shown in Figure 4.8, can be used to visually observe the sharpness of predictions as well as extreme values. Percentiles provide information on data distribution. The nth percentile is the value where $n\%$ of the data have that value or lower (45). For a cumulative density function, $F(x)$, the nth percentile is the value of x when $F(x) = n$. Figure 4.9 illustrates the CDF of one of the nine predictions in the percentile diagram 4.8. The 80th, 95th, and 100th percentile are highlighted in this CDF plot. As highlighted, the difference between the 95th and the 100th percentile is large compared to the difference between the 80th and

95th. This is due to extreme values of predicted ENS with very low probabilities. The flatness of the CDF curve at the highest ENS values give relatively wide gaps between percentile bands in the percentile diagram. In Figure 4.8 the gap between the 95th and 100th percentile is larger than gaps between the rest of the percentile gaps. This enables visual detection of outliers and extreme values.

In this study, percentiles from resulting CDFs are plotted in the y-axis with prediction periods on the x-axis, as shown in Figure 4.8. This is to be able to visually compare the sharpness of the predicted CDFs and also locate extreme values.

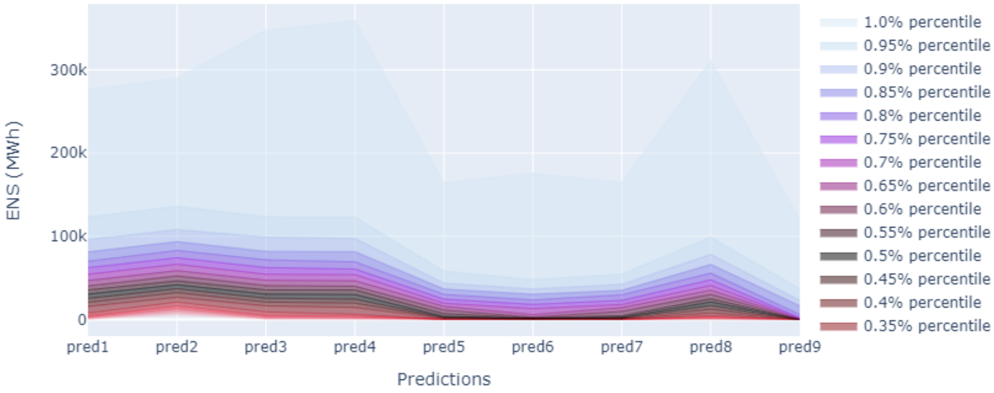


Figure 4.8: *Example of a percentile diagram. Percentile diagrams enable visual assessment and comparison of multiple predictions.*

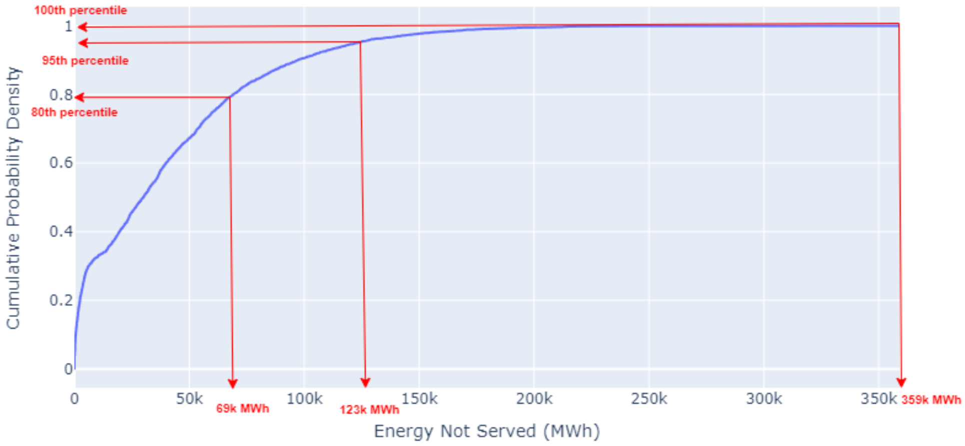


Figure 4.9: *Cumulative density plot of pred3 in Figure 4.8 with 80th, 95th and 100th percentile highlighted.*

.

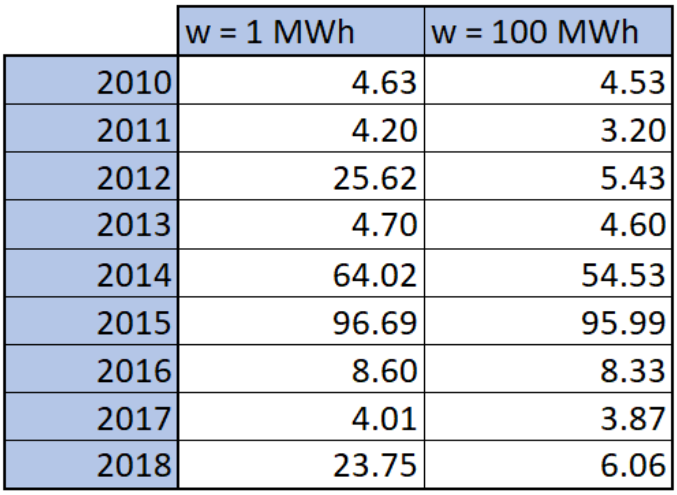
5 Results and Discussion

In this section, the findings of the research is presented and discussed. In the first subsection, the chosen verification methods for evaluating results from MONSTER will be discussed. In subsection 5.2 the results from MONSTER's accuracy for different time intervals will be presented and discussed. Subsection 5.3 will go through the sensitivity analysis and the changes for predicted ENS with changed input.

5.1 Evaluation of Methods for Verification

Methods chosen for this analysis are the CRPS score, reliability diagrams, and percentile diagrams. When computing the CRPS scores, dividing the raw data from the predictions into bins affect the resolution of the CDF, as discussed in the method section. The optimal bin width is the smallest possible, however, depending on the size of data, computing power could be a limitation. Due to the sample size of this study, computing power is not a problem, therefore the smallest possible width is chosen. Table 5.1 shows the CRPS score computed for each year in the yearly predictions, using two different bin widths. As illustrated in the table, most CRPS scores are not significantly affected by bin width. However, in years like 2012 and 2018 the CRPS score is significantly different for $w = 1MWh$ compared to $w = 100MWh$. This is because some of the variations in the CDF disappears when the data is generalised in a larger bin width. Using a a larger bin width – and therefore lower CDF resolution – can affect the end results remarkably. This should be taken into account when choosing bin width.

Table 5.1: *This table lists CRPS scores per instance for bin widths,* $w_1 = 1MWhandw_2 = 100MWh, foryearlypredictions.$



	w = 1 MWh	w = 100 MWh
2010	4.63	4.53
2011	4.20	3.20
2012	25.62	5.43
2013	4.70	4.60
2014	64.02	54.53
2015	96.69	95.99
2016	8.60	8.33
2017	4.01	3.87
2018	23.75	6.06

Like MAE, lower CRPS values represent a higher accuracy. Therefore, when comparing scores, lower values imply a more accurate result. Normalising the score gives a better understanding of the level of deviation from the actual value. However, what makes the value of a CRPS score good is difficult to define. This depends on the intended use of the model and the level of accuracy required.

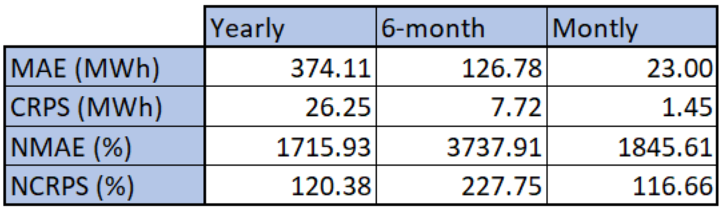
Reliability and percentile diagrams are used as complementary visual assessment tools. Since the data set is small, the reliability diagram is not expected to follow a diagonal line, however it can give valuable insights.

5.2 Prediction Accuracy

5.2.1 CRPS and MAE of Time Intervals

Table 5.2 shows the MAE, CRPS, NMAE, and NCRPS values for yearly, 6-month and monthly intervals. The NMAE is lower for year-intervals than for monthly and 6-month intervals. This suggests that the expected values of ENS are more accurate for yearly intervals. The NCRPS is lower for monthly time intervals than yearly, which indicates that the ENS-predictions for lower time intervals are more accurate. This was not expected, since MONSTER is suited for long-term reliability assessment and analysing the mean ENS in an area over a longer period of time should give a more accurate prediction. Possible reasons for the lower NCRPS score for monthly predictions will be discussed in subsection 5.2.2.

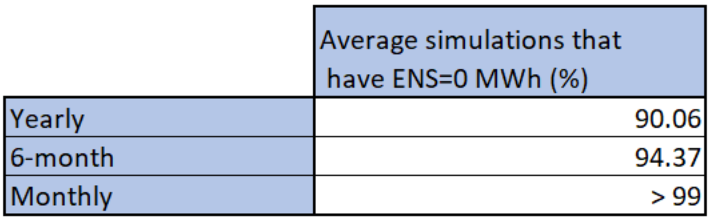
Table 5.2: *MAE, CRPS, NMAE and NCRPS of predictions for different time intervals.*



	Yearly	6-month	Montly
MAE (MWh)	374.11	126.78	23.00
CRPS (MWh)	26.25	7.72	1.45
NMAE (%)	1715.93	3737.91	1845.61
NCRPS (%)	120.38	227.75	116.66

Looking more closely at the predicted CDFs from MONSTER, the cumulative frequency is high for $ENS = 0MWh$ for all instances. As listed in Table 5.3, the average simulations that result in $ENS = 0MWh$ is 90.06 % for yearly intervals, 94.37 % for 6-month intervals and less than 1 % for monthly intervals. Figures 5.1 and 5.2, demonstrate the predicted and observed CDFs for an arbitrary yearly interval. In this specific instance, over 90 % of simulations do not have any predicted interruptions. This can be deducted from the CDF where the cumulative frequency at $ENS = 0MWh$ is above 90 %. Since CRPS is a score based on how much the predicted CDF deviates from the observed CDF, instances where the predicted CDF has a high cumulative frequency value for $ENS = 0MWh$ and an observed value of $ENS = 0MWh$ may score higher on accuracy, not necessarily because the results are more accurate, but because of the shape of the CDF. Monthly intervals have a larger number of observed instances with $ENS = 0MWh$ and this may contribute to the low CRPS score for monthly intervals, although they are not necessarily more accurate than yearly interval predictions. Subsection 5.2.2 will look more closely at predicted failures and interruptions to further explain the shape of the CDFs in this study.

Table 5.3: *Average percentage of simulations per prediction type with ENS=0 MWh.*



	Average simulations that have ENS=0 MWh (%)
Yearly	90.06
6-month	94.37
Monthly	> 99

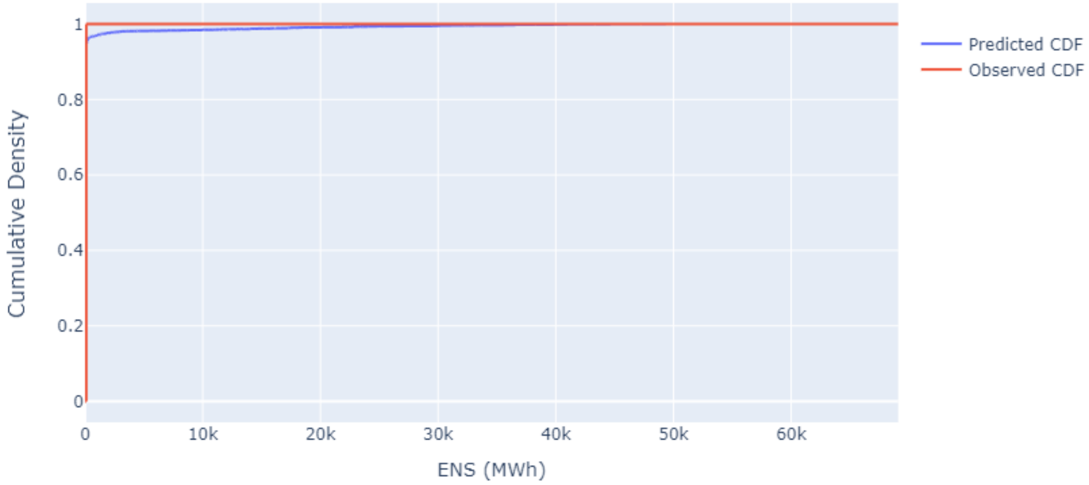


Figure 5.1: *Cumulative density plot for an arbitrary prediction and the historically observed value.*

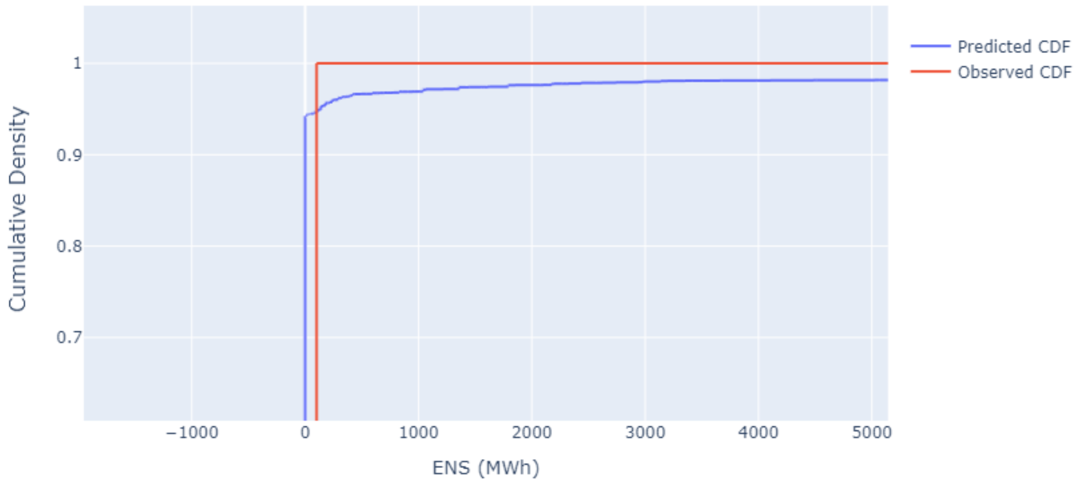


Figure 5.2: *Zoomed cumulative density plot for an arbitrary prediction and the observed value.*

5.2.2 The CDF and Probability

Assuming that the CDF represents the cumulative probability of ENS, for a perfectly reliable prediction, the mean cumulative probability of a given ENS-value should be equal to the mean observed cumulative frequency. Provided with enough data, the predicted cumulative probability and the observed cumulative frequency should be equal, for a perfect prediction. Since MONSTER's current use is long-term reliability it is unsure if enough data points can be provided. Even so, the reliability diagram can still give important insights to assess reliability in the predictions.

Figure 5.3 illustrates a reliability diagram for yearly intervals. The horizontal line up to $x = 0.57$ is due to nominal levels to that point having $ENS = 0MWh$. The observed curve crosses the y-axis at $y = 0.33$, which means the observed frequency of $ENS = 0MWh$ is 33 %. The predicted probability of not having any interruptions is higher than 33 %. Also, the reliability function is steep from $x = 0.93$ to $x = 97$. This indicates that the majority of the observed yearly ENS-values fall within a 4 % portion of the predicted ENS-values. This further confirms that predictions for this analysis contain a high number of simulations that lead to $ENS = 0MWh$ compared to observed values and the highest simulated years with $ENS \neq 0MWh$ have significantly higher ENS-values than observed.

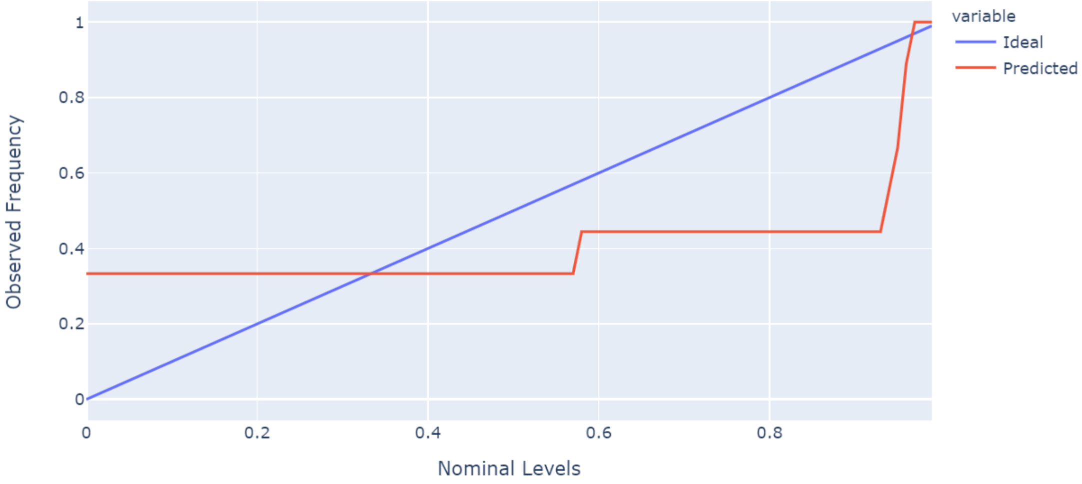


Figure 5.3: *Reliability diagram for yearly ENS-predictions. The diagonal line is the ideal.*

To look closer at interruptions simulated by MONSTER, duration, and power have been retrieved from the interruptions for yearly intervals. These are compared with power and duration of observed values. Observed interruptions from 1998 to 2018 are used as well as a random sample of 200 interruptions from the yearly predictions. Figure 5.4 presents the power and duration of interruptions in predicted and observed data. The predicted and observed interruptions seem to have the same power range. However, the durations predicted by MONSTER are much higher than the observed durations. The high predicted durations will lead to a higher ENS per predicted interruption in MONSTER compared to historical ENS. This may contribute to the long tail in the CDF.

To inspect more closely, all interruptions from the yearly predictions are plotted in Figure 5.5. Since there are over 100 000 points, the scatter plot is made with an opacity to clarify areas with a higher concentration of points. As illustrated in the figure, most points are concentrated in lower durations. The predicted durations, however, are still significantly higher than observed durations. There is a high concentration of points around 400 hour and 500 hour durations. This is due to discrete failure durations, which will be discussed in later sections.

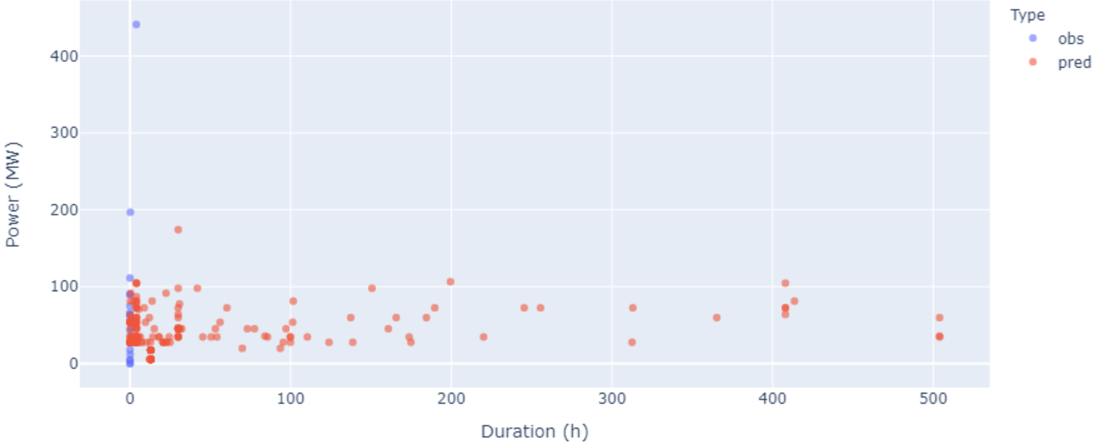


Figure 5.4: *Scatter plot of power vs. duration for predicted and observed failures. This is for all observed failures in the historical data for the area of analysis and 200 randomly selected samples from the yearly predictions from 2010 to 2018.*

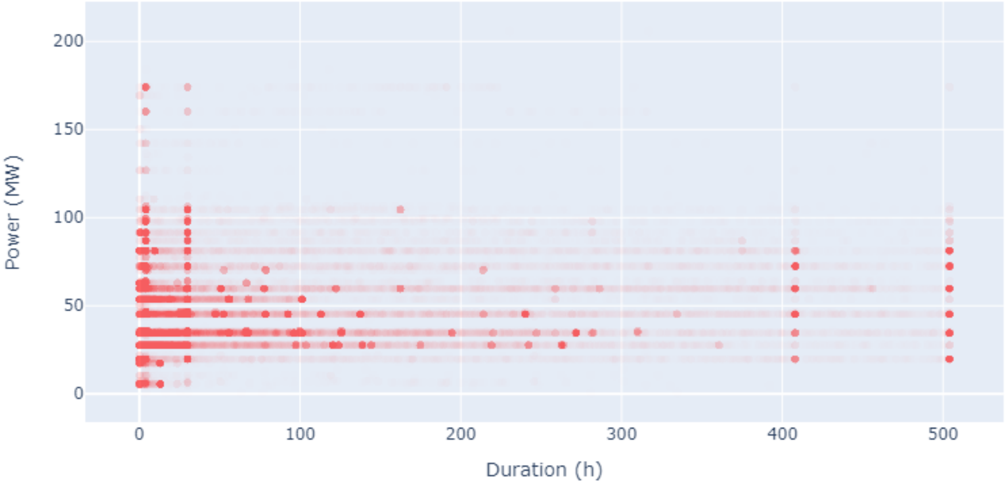
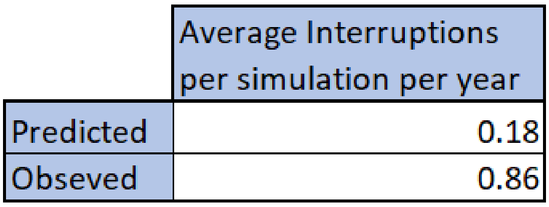


Figure 5.5: *Scatter plot of power vs. duration for all predicted interruptions in the yearly predictions. The plot has opacity to clarify areas with high concentration of points.*

.

Table 5.4: *Average interruptions per year per simulation for all yearly predictions and average interruptions per year from historical data.*



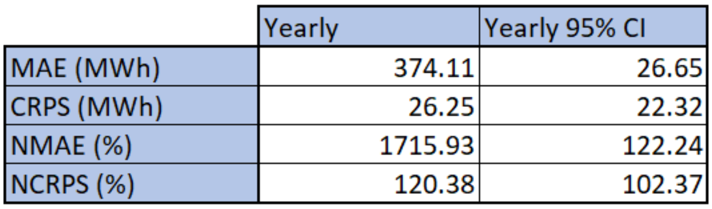
	Average Interruptions per simulation per year
Predicted	0.18
Obseved	0.86

5.2.3 Extreme Values

Although the extreme values are important for risk analysis, they do significantly affect the expected value of ENS predicted by MONSTER. A suggested way of decreasing this effect and acquiring more accurate predictions is by using an interval of the predicted CDF instead of the full density. Table 5.5 illustrates the MAE and CRPS of yearly intervals with full density and the a 95 % confidence interval. Both MAE and CRPS are lower for the 95 % confidence interval results. However, with the shape of CDF in the results presented, filtering out the top 2.5 % of the full CDF will remove out a large amount of the CDF resulting $ENS > 0MWh$. Based on the results of NMAE and NCRPS the accuracy may seem to have increased for these specific predictions. Even so, using an interval prediction with 95 % confidence interval for monthly predictions, may filter out all simulations where $ENS > 0MWh$ for some predictions. For example, for monthly predictions – where less than 1 % of the simulations result in $ENS \neq 0$ – a 95 % confidence interval would remove all simulations where $ENS \neq 0$. Therefore, using interval predictions require an assessment of the CDF and detection of extreme values.

There is, however, a significant reduction in the NMAE from full density prediction to the 95 % confidence interval prediction. This further strengthens the assumption that extreme values highly affect the predicted expected value, and therefore the importance of their detection.

Table 5.5: *MAE, CRPS, NMAE and NCRPS for full density prediction and 95 % central interval prediction.*



	Yearly	Yearly 95% CI
MAE (MWh)	374.11	26.65
CRPS (MWh)	26.25	22.32
NMAE (%)	1715.93	122.24
NCRPS (%)	120.38	102.37

5.3 Sensitivity Analysis

This subsection goes through the results and discussion for the sensitivity analysis. The base case for the sensitivity analysis is yearly intervals including all inputs. Sensitivity analysis were performed on yearly predictions by excluding different inputs and changing the failure rate for station components.

5.3.1 Analysis with only Lines and Cables

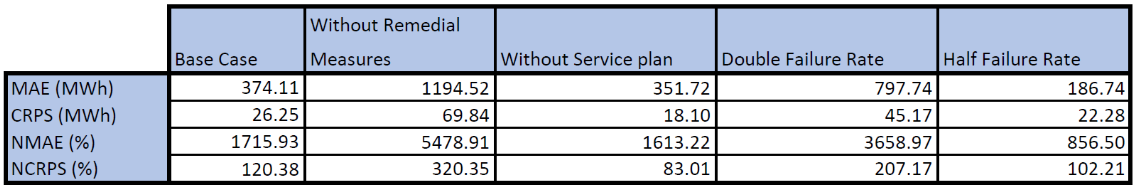
Predictions including lines and cables and excluding station components were made for 2010 and 2011. They did not result in any interruptions and therefore did not compute ENS predictions. This is because the failure rates for lines and cables in the area of analysis are low. Because of this, no more runs were made of this type. MONSTER runs were also made for only lines, and cables, excluding remedial measures. This is because it is more likely that an analysis of only lines and cables will have interruptions if remedial measures are excluded. These predictions did either not result in any interruptions or did not converge. Therefore, analysis with lines and cables are excluded from the result tables and figures. This is a great indication that failures in lines and cables may not contribute much to the results of this analysis.

5.3.2 Without Remedial Measures

Looking at Table 5.6 and Figure 5.6, it is clear that the remedial measures input is essential for the predicted ENS. The MAE and CRPS are significantly higher for predictions without remedial measures compared to the base case. Figure 5.6 presents the expected value of ENS predicted each year in the analysis. The expected ENS values without remedial measures are 3 times higher than the expected ENS values for the base case, as shown in Table 5.8. The lack of remedial measures in the model leads to more of the simulated contingencies resulting in interruptions. As seen in Table 5.7 average interruptions per year are significantly higher for MONSTER-runs without remedial measures than for the base case. Additionally, excluding remedial measures as an input in the predictions does not correlate with reality, since the System Operations Center operates the power system and continuously takes measures to prevent and reduce the consequences of interruptions. Not including remedial measures overlooks this work, and therefore an essential factor that affects the transmission system's reliability.

The results for predictions without remedial measures compared to the base case indicate that this input has a notable effect. Table 5.8 depicts a 207.23 % increase in the average expected ENS. This raises question on to what degree simplification of the input for this study affects the accuracy of predictions. The system operation centre operates using multiple remedial measures. Also, the move load-measure that is added to the MONSTER-predictions allows 100 % of a load to be transferred to the neighbouring electrical substation, regardless of the system state. This is a simplification that may lead to more interruptions being prevented in the model compared to what is realistically possible. For example, during winter – since the power loads are higher – the system capacity will be low compared to summer.

Table 5.6: *MAE, CRPS, NMAE and NCRPS for sensitivity analysis.*



	Base Case	Without Remedial Measures	Without Service plan	Double Failure Rate	Half Failure Rate
MAE (MWh)	374.11	1194.52	351.72	797.74	186.74
CRPS (MWh)	26.25	69.84	18.10	45.17	22.28
NMAE (%)	1715.93	5478.91	1613.22	3658.97	856.50
NCRPS (%)	120.38	320.35	83.01	207.17	102.21

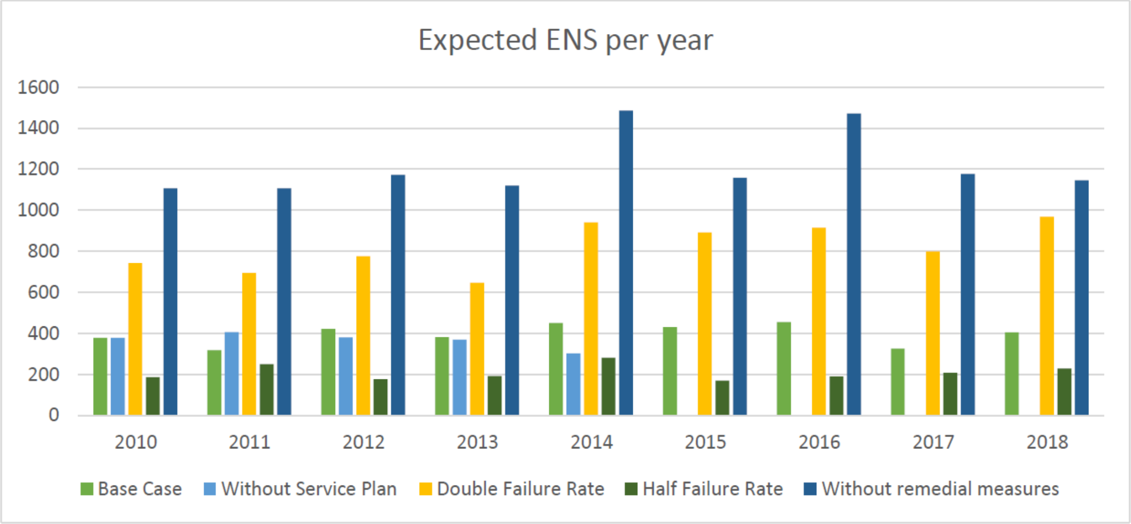
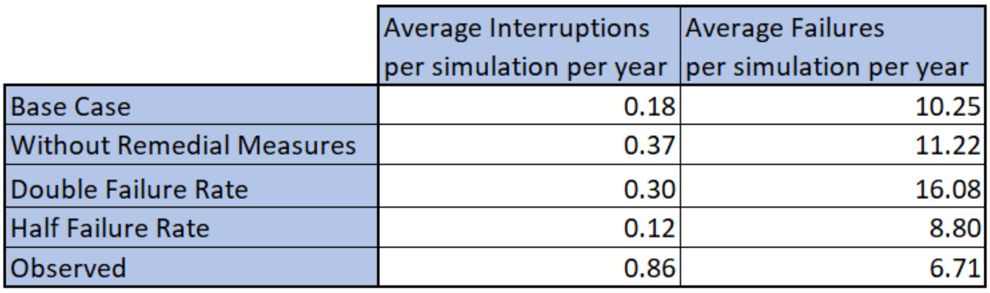


Figure 5.6: *Expected value of ENS for predictions from sensitivity analysis.*

Table 5.7: *Average predicted failures and interruptions per simulation per year for predictions and historically observed values.*

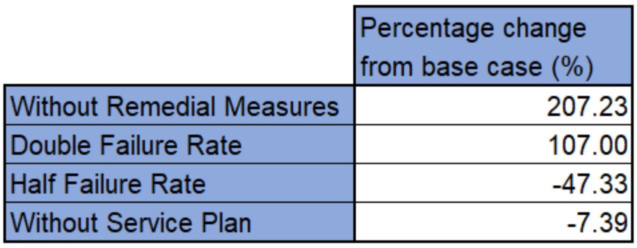


	Average Interruptions per simulation per year	Average Failures per simulation per year
Base Case	0.18	10.25
Without Remedial Measures	0.37	11.22
Double Failure Rate	0.30	16.08
Half Failure Rate	0.12	8.80
Observed	0.86	6.71

5.3.3 Failure Rates

Changing the failure rates – and therefore the failure probabilities – affect the number of contingencies in the simulation period. In this part of the sensitivity analysis, failure rates for station components were doubled and halved to analyse the effect on the predicted ENS. As seen in Table 5.8 the expected ENS is roughly twice as high for increased failure rates and a half for the halved failure rates. Since the failure rates for lines and cables in the Greater Oslo Region are low, most of the resulting ENS from these predictions are presumably based on failures in station components. Therefore this correlation between predicted ENS and change in failure rates could indicate a linear association. A more in-depth analysis of failure rates could reveal if there is a linear association. However, looking at the average number of failures on Table 5.7 there is no indication that the number of simulated failures doubled or halved with the failure rate. Nevertheless, as Table 5.7 illustrates, there is an increase in average interruptions and failures for predictions with a double failure rate and a decrease for predictions with half failure rates.

Table 5.8: *Percentage change of average predicted ENS from base case.*



	Percentage change from base case (%)
Without Remedial Measures	207.23
Double Failure Rate	107.00
Half Failure Rate	-47.33
Without Service Plan	-7.39

5.3.4 Service Plan

MONSTER-runs without a service plan were made for only 5 years. Figure 5.7 plots the expected values from predictions without a service plan and the base case. As seen here, a lack of service plan generally decreases the expected ENS value. Adding planned maintenance as an input in the model makes components unavailable for the service period. The transmission system is therefore more sensitive in case of a failure. Therefore, lower ENS is expected for predictions without a service plan input. The effect of the service plan input is not as significant as remedial measures. As listed in Table 5.8, the percentage difference in predicted expected ENS from the base case is 7.39 %. This is much smaller than for predictions without remedial measures and for changed failure rates.

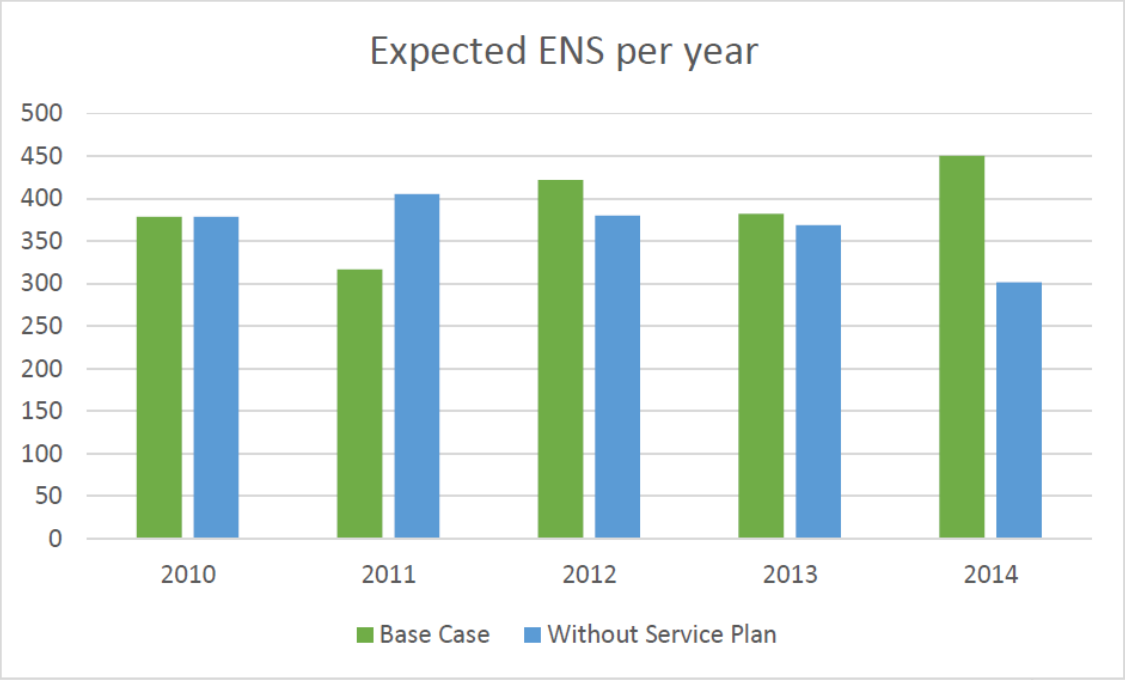


Figure 5.7: *Expected value of ENS for base case and predictions without a service plan.*

5.4 Analysis of Percentile Diagrams

5.4.1 Extreme Values

Percentile diagrams provide information on CDF distributions. The gap between band widths can also be a useful tool to detect extreme values. The percentile diagram for predictions without remedial measures in Figure 5.8 illustrates a relatively high gap between bands clearly. The top 2.5 % of values are much higher than the rest of the percentile bands. The base case too shows a higher band hap between the 100th and 99.5th percentile in Figure 5.10. When excluding the top 2.5 % of the CDF, by using the 95 % confidence interval, the years 2010 to 2013, as seen in Figure 5.11, do not have any percentiles $\hat{q}^n > 0$. Assuming the CDF represents probability, using an interval prediction of 95 % will result in there being no probability of interruptions for these years. Although the results indicate that an interval prediction is more accurate, in subsection 5.2.3, these results are not informative.

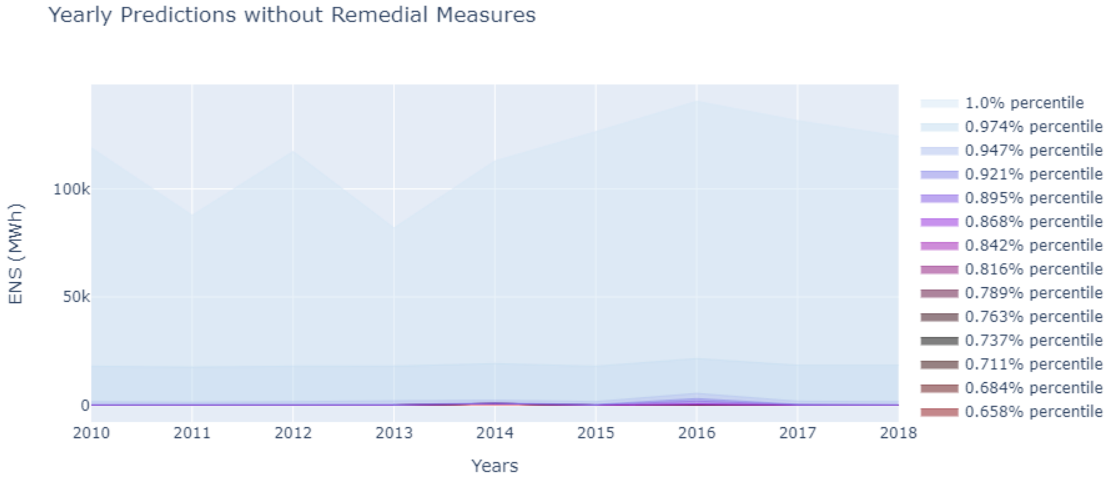


Figure 5.8: *Percentile diagram for predictions without remedial measures, to 100th percentile.*

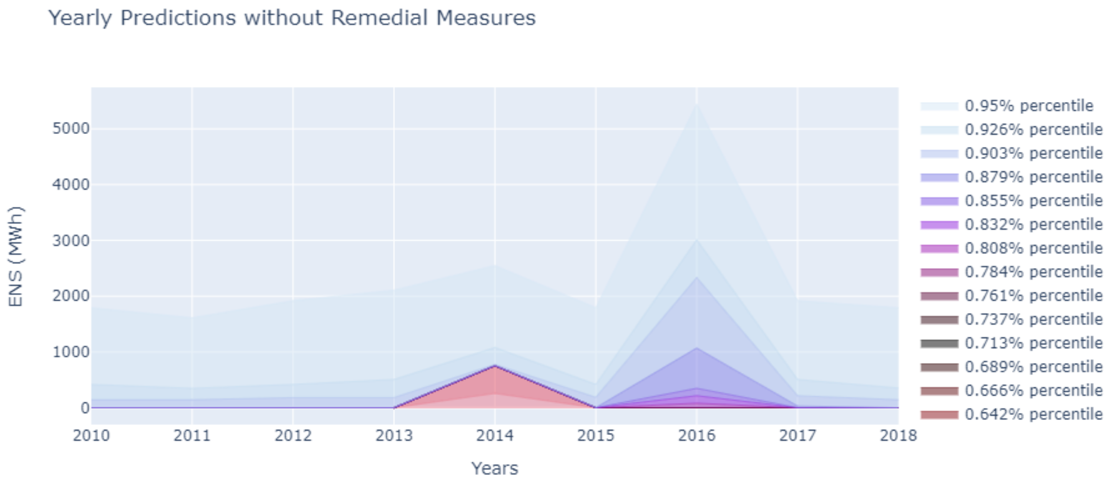


Figure 5.9: *Percentile diagram for predictions without remedial measures, to 95th percentile.*

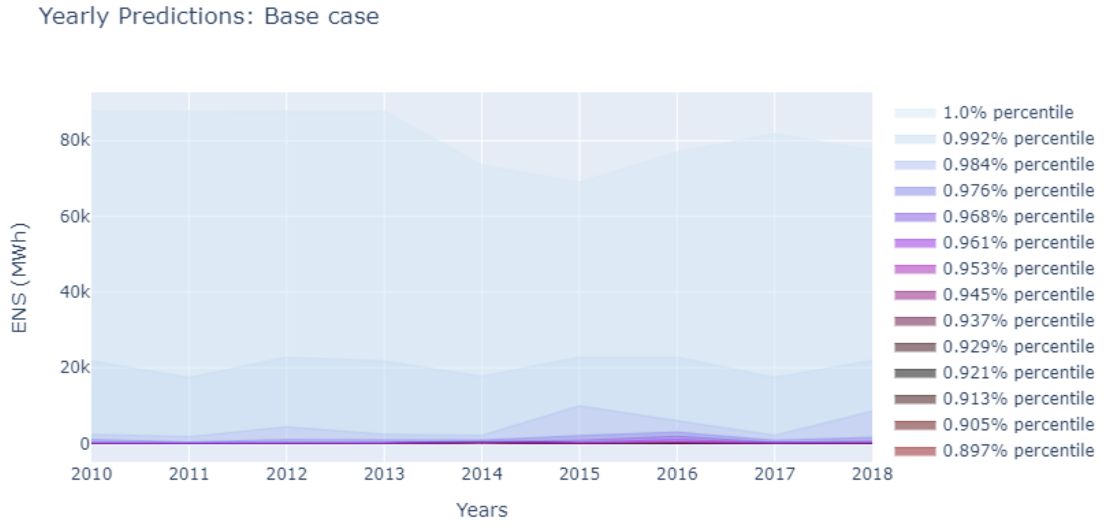


Figure 5.10: *Percentile diagram for base case, to 100th percentile.*

5.4.2 Remedial Measures

The percentile diagram for predictions without remedial measures show a higher number of percentiles being above $ENS = 0MWh$ than for the base case. The highest percentiles also have higher ENS values. This illustrates that simulated failures without remedial measures lead to significantly higher predicted ENS. Excluding remedial measures in the predictions overlooks the System Operation Center's work by taking measures to prevent power outages. These results complement the earlier

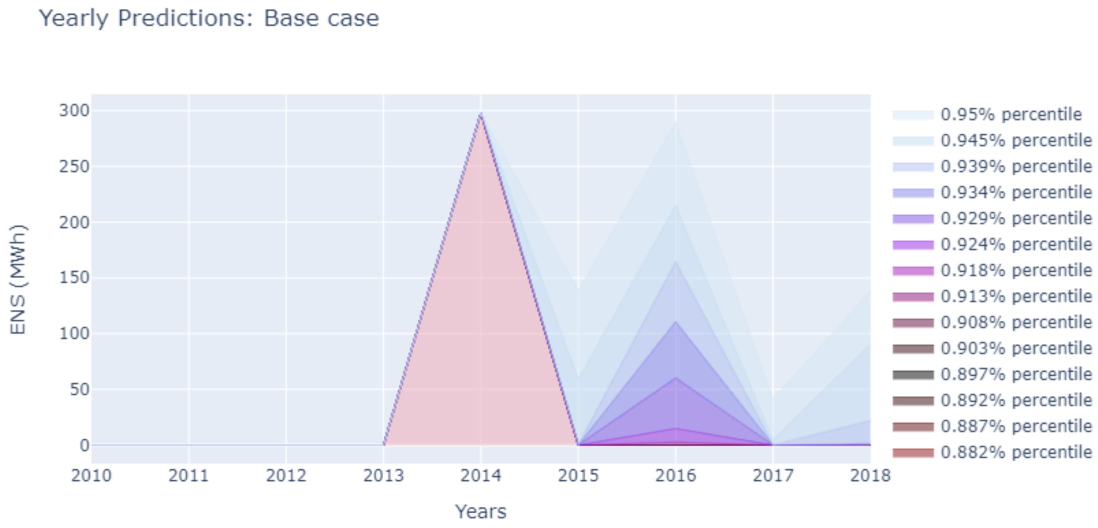


Figure 5.11: *Percentile diagram for base case, to 95th percentile.*

presented average interruptions per year for predictions without remedial measures compared to the base case. The average expected ENS is also around 3 times as high for predictions without remedial measures. The simulation tool's sensitivity to the remedial measure input is evident.

The simplifications made when adding remedial measures may lead to more interruptions being prevented than is feasible. The move load-measure allows 100 % of the load to be moved to the neighbouring transformation station. It does not consider limited capacity in the regional transformation system due to, for example, high loads. Since this input has the largest effect on the results from MONSTER, an expansion of this input, and possibly adding multiple remedial measures, could increase accuracy.

5.4.3 Variations in Yearly Prediction

The percentile diagram for the base case in Figure 5.10 does not show much variation between the predicted ENS per year. For this study, the inputs that have been changed are service plans and weather. MONSTER has weather dependent failure rates only for lines and not for cables, and station components. This is because these components are not as exposed to weather changes as lines. In the area of analysis for this study, the failure rates for lines are low and therefore insignificant compared to failure rates for other components. Weather is therefore not expected to have a notable effect on the predicted ENS. This can explain some of the lack of variation in the predicted ENS.

However, there is a notable difference in the year 2014, which peaks in the percentile diagrams for the base case and predictions without remedial measures. 2014 is the only simulated year that has more than 40 % of the simulations leading to $ENS > 0$. The year 2016 also peaks, although not as much as 2014. These peaks are most probably caused by the service plans input. Looking at the percentile diagram for predictions without service plans in Figure 5.12 there is no peak in 2014. Also, the plot of expected values in Figure 5.7, the difference between the predicted ENS for the base case and the predictions without a service plan were highest in 2014. This indicates some effect of changing the service plan input on ENS results.

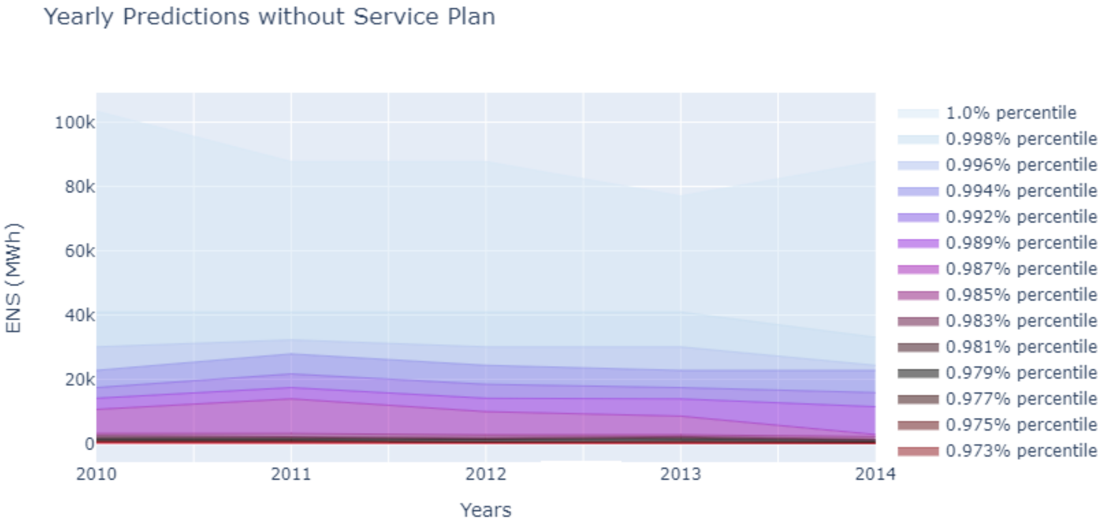


Figure 5.12: *Percentile diagram for predictions without a service plan.*

5.4.4 6-month and monthly intervals

6-month predictions and monthly predictions have much fewer simulations that lead to interruptions. For both predictions, summer months have more simulations with $ENS \neq 0$. For the 6-month predictions, winter months have higher ENS for the highest percentiles. Since the load is higher during the winter it is expected that interruptions in this period will have a higher ENS. There are also significantly higher numbers of service plans during the summer months compared to the winter months. This can explain the higher number of predicted interruptions during summer periods.

The peaks, in the percentile diagrams, for 6-month and monthly predictions were analysed to inspect for any correlations that might exists between these pinnacles, and the periods with historically observed interruptions. Some trends were observed for 6-month predictions. No trends were found for monthly predictions.

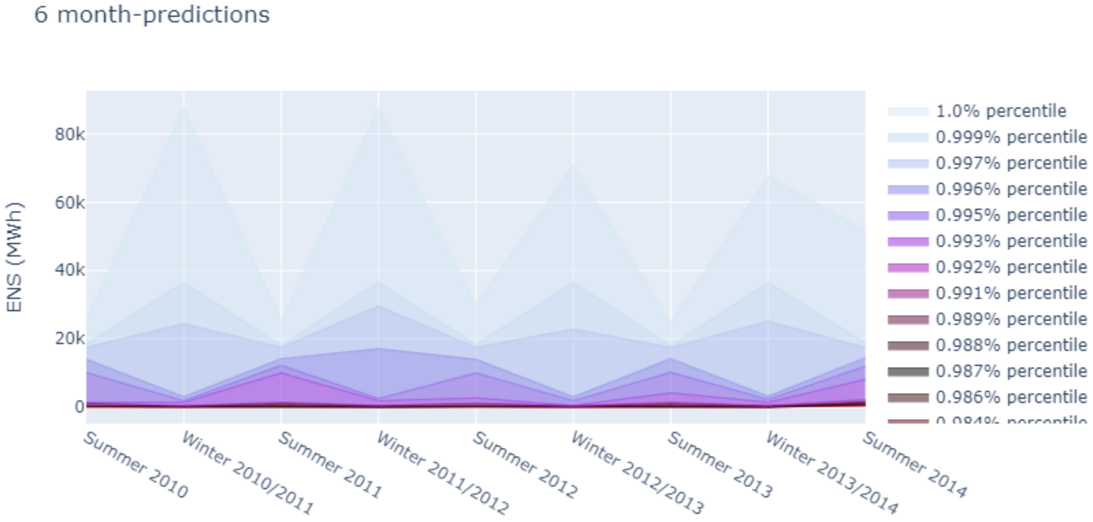


Figure 5.13: *Percentile diagram for 6-month intervals.*

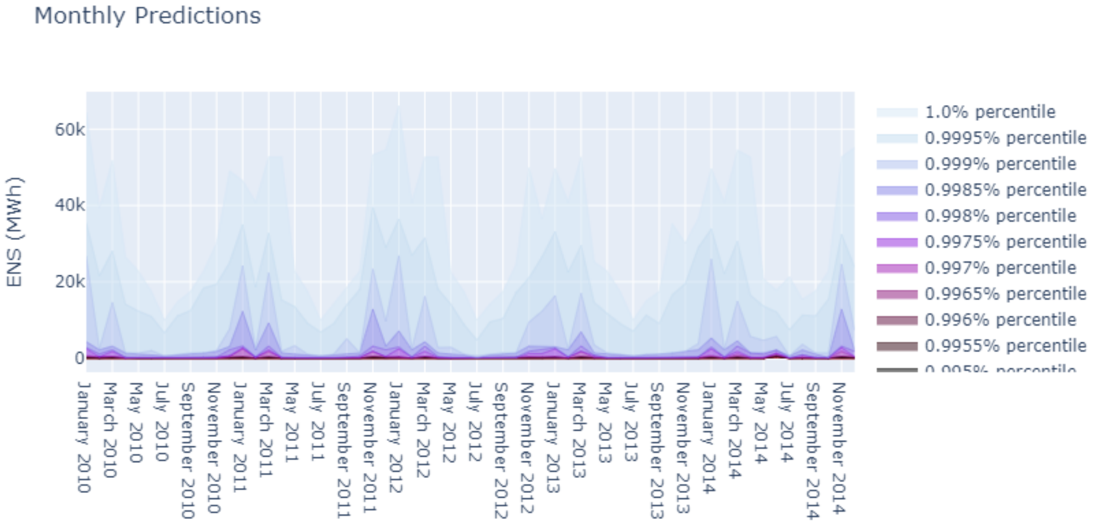


Figure 5.14: *Percentile diagram for month intervals.*

5.5 General Discussion

5.5.1 The Cumulative Density Function

The results presented indicate that interruptions are not simulated in the tool as often as historically observed. For yearly predictions, except 2014, all other years in the analysis have $ENS = 0MWh$ for more than 90 % of the simulated years. There are also some simulations resulting in high ENS values. This leaves the CDF very steep at low ENS values and flat at high ENS values. A possible explanation for this is the failure durations drawn in MONSTER for station components. Since the failure durations are discrete for station components and cables, failures in these components may result in either low or very high predicted ENS. An other explanation for the steep CDF at low ENS values is redundancy in the power system for the area of analysis, therefore few contingencies lead to interruptions. Table 5.7 show that average simulated failures in the yearly predictions are much higher than the average predicted interruptions. This is also the case for predictions without added remedial measures. This indicates that most simulated failures do not lead to interruptions in the simulation tool. Since there are no remedial measures added to these predictions, this further strengthens the assumption of redundancy.

5.5.2 Time Span of Predictions

Considering that MONSTER is currently not suited for short term analysis, the time intervals of the predictions in this analysis is a disadvantage when assessing the accuracy of the tool. As discussed in previous sections, MONSTER is a simulation tool for assessing long-term reliability. However, the purpose of using shorter time spans in this study was also to evaluate the tool's ability to predict accurately for short term analysis. The overall results indicate a better performance for longer time intervals. A comparison between results from this study and long-term accuracy – of 10 years or more – requires a more in-depth study of multiple areas, since there is not enough historical data to conduct a long-term reliability accuracy assessment with one area of analysis.

5.5.3 Data Points

The main challenge of this study was the amount of data. MONSTER is currently not suited for predicting ENS for the short time periods of this research. Therefore, analysing the accuracy for smaller time intervals goes against the tool's intended use. On the contrary, using longer time intervals results in a smaller data set, and therefore not enough foundation to conclude on the tool's performance.

Analysing of a longer time period than 9 years or more areas would result in more work to set up MONSTER runs, which was limited by the scope of this study. Although yearly intervals over 9 years does not give enough grounds for a conclusion on the tool's performance, the results have given interesting insights and grounds for further research.

Also, a significant part of this study is the assessment of methods to use for verification of probabilistic simulation models for reliability in transmission systems in general and MONSTER in specific. This part of the research did not require much data and the results from the case study did provided insights to the chosen verification methods' limitations.

5.5.4 Simplifications and Limitations

Many simplifications have been made for this research, for instance, the same system states have been used for all runs. This simplification was made due to few changes in the system state in the period of analysis.

Individual and weather dependent failure probabilities are a very important feature in the simulation tool. This feature is added for lines, since overhead lines are more exposed to weather changes than station components and cables. Considering that the Greater Oslo Region has very low failure rates for lines, this study does not examine that feature. Therefore the accuracy assessment in this research may be biased.

The failure durations in Figure 5.5 illustrates a higher concentration of failures for lower durations. However there is a concentration of points at points $400h$ and $500h$. The randomly drawn failure durations for temporary failures for lines follow continuous distributions based on historical durations for each failure type. Because of the priority of modelling failure in lines and the simulation tool still being under development, the durations drawn for outages due to failure in transformers are discrete, hence outages will have specific values representing short or long outage durations. The permanent failures durations are also discrete (46).

5.5.5 N-1 Criterion

As discussed in the theory section, the N-1 criterion overlooks the probabilities of component failures, as well as the likelihood of multiple components failing. This can lead to an over-investment and therefore a probabilistic tool that accurately predicts reliability worth would be in the interest of the TSO in optimise socioeconomic costs. This study has tested the accuracy of the simulation tool in the Greater Oslo Region and there are some indications of higher predicted ENS for instances that have higher observed ENS, a more expanded research is required to make a general evaluation. Replacing the N-1 criterion is not necessarily the goal. However, probabilistic tools like MONSTER open up for a more flexible use of the N-1 criterion in transmission system planning. Since the simulation tool includes uncertainty of inputs, it can also recognise areas where higher reliability than N-1 is socioeconomically beneficial. Probabilistic tools do not necessarily have to replace the N-1 criterion but can complement and support in decision making when prioritising expansion plans in the transmission grid.

Disregarding whether probabilistic tools can serve as an alternative to the N-1 criterion, MONSTER provides the possibility of a holistic, informative, and effective evaluation of reliability worth, and is nonetheless valuable for future transmission planning.

5.5.6 Verification Methods

Although the chosen score method, CRPS, has worked well for this study, the low score for monthly intervals has revealed its challenges. Since CRPS averages over the whole range of thresholds and probabilities, it does not identify deficiencies in the predicted CDF. In *Decomposition and graphical portrayal of the quantile score*, Sabrina Bentzien and Petra Friederichs suggest using the Quantile Score (QS) for different probability levels to detect tails in the probability distribution (47).

In this study, reliability and percentile diagrams are used to visually inspect the distributions and detect extreme values. Other visual tools, such as the Rank Histogram and Probability Integral Transform (PIT) diagrams are also useful when detecting tails in the probability distribution (33). A more in-depth study of verification methods and the use of multiple assessment tools would be beneficial for future evaluations of MONSTER predictions.

6 Conclusion and Further Research

6.1 Conclusion

The main objective of this thesis was to assess methods for verification of probabilistic results. Methods from the weather forecasting community have been assessed and applied to evaluate the results from MONSTER. The score method, CRPS, was chosen to evaluate probabilistic ENS predictions from MONSTER. Complementary visual assessment tools have been used, which have given more insights for the results. It was discovered, by analysing percentile and reliability diagrams, that the percentage of simulated years with $ENS > 0MWh$ is very low compared to historically observed years with $ENS > 0MWh$. Furthermore, the yearly simulations with $ENS > 0$ have very high ENS-values. It was also discovered that CRPS overlooks deficiencies in the probability distribution.

The accuracies given by CRPS for predicted ENS were better for smaller time intervals. This does not necessarily denote that MONSTER predicts better for lower time intervals, but that the CDFs are closer to the observed values for monthly intervals. Looking at the shape of the CDFs, the cumulative frequency is high for $ENS = 0$, and there are more points with $ENS_{obs} = 0$ for monthly intervals. This may lead to a lower CRPS and NCRPS score for monthly intervals. It is suggested that including verification methods that detect deficiencies in the probability distribution will give better grounds for evaluating results from MONSTER. NMAE is lowest for yearly intervals, as expected. This indicates that the expected ENS values are more accurate for predictions with larger time spans.

The sensitivity analysis shows that this study is most sensitive to remedial measures. This does not mean that the tool in general is most sensitive to remedial measures. The results do show strong indications that this is an important input for the ENS significantly. It is suggested that an expansion of this input for further research will increase accuracy. Removing the service plan input did not have a notable effect on the results, for this study. It is also clear that in this area of analysis, lines and cables do not have as much effect on the predicted ENS as station components.

6.2 Further Research

The assessment of results from the case study have clarified the limitations of chosen verification methods. Further research of verification methods would be useful to ensure a holistic assessment of the probabilistic results from MONSTER. Since CRPS is not fit to detect extreme values, further research on verification methods is required. Additionally, using multiple scoring methods would give better grounds for evaluating the results from the simulation tool.

Since the main application for the simulation tool is long-term reliability, typically 10 to 40 years, assessing MONSTER's accuracy requires an analysis of multiple areas. An analysis of multiple areas can also be beneficial for developing a general method for handling the effect of extreme ENS values.

Furthermore, because of low failure rates in lines for Greater Oslo Region, individual, weather dependent failure rates have not been tested in this study. An assessment of MONSTER's performance in more areas can give more insights on this feature.

References

[1] Pourbeik P, Chakrabarti B, George T, Haddow J, Illian H, Nighot R, et al. Review of the current status of tools and techniques for risk-based and probabilistic planning in power systems. CIGRE. 2010;.

[2] Li W. Probabilistic transmission system planning. vol. 65. John Wiley & Sons; 2011.

[3] Agency for the Cooperation of Energy Regulators. Decision No 07/2019 of the Agency for the Cooperation of Energy Regulators. Agency for the Cooperation of Energy Regulators; 2019.

[4] ENTSO-E. Workshop "Data Collection for Probabilistic Risk Assessment"; 2020 (accessed December 10th, 2020). https://www.entsoe.eu/events/2020/01/17/workshop-data-collection-for-probabilistic-risk-assessment/.

[5] North-American Electric Reliability Corporation. Probabilistic Adequacy and Measures Technical Reference Report. North-American Electric Reliability Corporation; 2018 (accessed December 10th, 2020).

[6] Bronmo G, Vergnol A, Gamov N, Bukhsh W, Bell K, Andreev A, et al. Generally Accepted Reliability Principle with Uncertainty Modelling and Through Probabilistic Risk Assessment: Functional Analysis of System Development Process. 2015;.

[7] Statnett SF. Et helelektrisk Norge er innen rekkevidde; 2018 (accessed December 10th, 2020). https://www.statnett.no/om-statnett/nyheter-og-pressemeldinger/Nyhetsarkiv-2018/Et-helelektrisk-Norge-er-innen-rekkevidde/.

[8] Statnett SF. Konkurransekraft er nødvendig for å lykkes med elektrifisering; 2019 (accessed December 10th, 2020). https://www.statnett.no/om-statnett/nyheter-og-pressemeldinger/nyhetsarkiv-2019/konkurransekraft-er-nodvendig-for-a-lykkes-med-elektrifisering/.

[9] DNV GL. Strømnettet i et fullelektrisk Norge. DNV GL; 2019 (accessed December 10th, 2020). https://www.energinorge.no/rapport-stromnett.

[10] The Norwegian Water Resources and Energy Directorate. Kraftåret 2017: Rekordhøyt strømforbruk og høye priser til tross for vått og varmt år; 2017 (accessed December 10th, 2020). https://www.nve.no/nytt-fra-nve/nyheter-energi/kraftaret-2017-rekordhoyt-stromforbruk-og-hoye-priser-til-tross-for-vatt-og-varmt-ar/.

[11] Bruvik K, Hytten LM. Probabilistic Reliability Analysis in the Norwegian Transmission System. In: 2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS). IEEE; 2020. p. 1–6.

[12] The Norwegian Ministry of Petroleum and Energy. Fakta - Energi og Vannressurser i Norge; 2012 (accessed December 15th, 2020). https://www.regjeringen.no/globalassets/upload/oed/faktaheftet/fakta_energi_og_vannressurs.pdf.

[13] Norwegian Water Resources and Energy Directorate. The Norwegian power system. Grid connection and licensing; 2018 (accessed December 10th, 2020). http://publikasjoner.nve.no/faktaark/2018/faktaark2018_03.pdf.

[14] The Norwegian Ministry of Petroleum and Energy. Strømnettet; 2019 (accessed December 10th, 2020). https://energifaktanorge.no/norsk-energiforsyning/kraftnett/.

[15] Statnett SF. Årsrapport 2019. Statnett SF; 2019 (accessed December 10th, 2020). https://www.statnett.no/om-statnett/nyheter-og-pressemeldinger/nyhetsarkiv-2020/arsrapport-2019/.

[16] The Norwegian Ministry of Petroleum and Energy. Forskrift om leveringskvalitet i kraftsystemet; 2004 (accessed December 10th, 2020). https://lovdata.no/dokument/LTI/forskrift/2004-11-30-1557.

[17] The Norwegian Ministry of Petroleum and Energy. Forskrift om systemansvaret i kraftsystemet; 2002 (accessed December 10th, 2020). https://lovdata.no/dokument/LTI/forskrift/2004-11-30-1557.

[18] Vrana TK, Johansson E. Overview of power system reliability assessment techniques. CIGRE 2011. 2011;p. 51–62.

[19] The Norwegian Water Resources and Energy Directorate. KILE – kvalitetsjusterte inntektsrammer ved ikke-levert energi; 2009 (accessed December 14th, 2020). https://www.nve.no/reguleringsmyndigheten/okonomisk-regulering-av-nettselskap/om-den-okonomiske-reguleringen/kile-kvalitetsjusterte-inntektsrammer-ved-ikke-levert-energi/.

[20] Bronmo G, Vergnol A, Gamov N, Bukhsh W, Bell K, Andreev A, et al. Generally Accepted Reliability Principle with Uncertainty Modelling and Through Probabilistic Risk Assessment: Functional Analysis of System Development Process. 2015;.

[21] Uusitalo L, Lehikoinen A, Helle I, Myrberg K. An overview of methods to evaluate uncertainty of deterministic models in decision support. Environmental Modelling & Software. 2015;63:24–31.

[22] Morales JM, Conejo AJ, Madsen H, Pinson P, Zugno M. Renewable energy sources—Modeling and forecasting. In: Integrating Renewables in Electricity Markets. Springer; 2014. p. 15–56.

[23] Statnett SF. Vi fortsetter elektrifiseringen av Norge; 2020 (accessed December 10th, 2020). https://www.statnett.no/om-statnett/nyheter-og-pressemeldinger/nyhetsarkiv-2020/vi-fortsetter-elektrifiseringen-av-norge/.

[24] Solheim ØR, Warland L, Trötscher T. A holistic simulation tool for long-term probabilistic power system reliability analysis. In: 2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS). IEEE; 2018. p. 1–6.

[25] Kroese DP, Brereton T, Taimre T, Botev ZI. Why the Monte Carlo method is so important today. Wiley Interdisciplinary Reviews: Computational Statistics. 2014;6(6):386–392.

[26] Cho WKT, Liu YY. Sampling from complicated and unknown distributions: Monte Carlo and Markov Chain Monte Carlo methods for redistricting. Physica A: Statistical Mechanics and its Applications. 2018;506:170–178.

[27] Ata MY. A convergence criterion for the Monte Carlo estimates. Simulation Modelling Practice and Theory. 2007;15(3):237–246.

[28] Box GE, Tiao GC. 1. In: Bayesian inference in statistical analysis. John Wiley Sons, Ltd; 1992. p. 1–75.

[29] Solheim ØR, Trötscher T, Kjølle G. Wind dependent failure rates for overhead transmission lines using reanalysis data and a Bayesian updating scheme. In: 2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS). IEEE; 2016. p. 1–7.

[30] Solheim Specialist in the Data Science Department, Statnett SF; Otober 7th 2020. Personal communication.

[31] Hammersley DC J M ; Handscomb. Monte Carlo Methods. Methuen; 1964.

[32] Pinson P, Nielsen HA, Møller JK, Madsen H, Kariniotakis GN. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology. 2007;10(6):497–516.

[33] Lauret P, David M, Pinson P. Verification of solar irradiance probabilistic forecasts. Solar Energy. 2019;194:254–271.

[34] Tastu J. Short-term wind power forecasting: probabilistic and space-time aspects. 2013;.

[35] Robertson A, Vitart F. Sub-seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting. Elsevier; 2018.

[36] Wilks DS. Statistical methods in the atmospheric sciences. vol. 100. Academic press; 2011.

[37] Hamill TM. Reliability diagrams for multicategory probabilistic forecasts. Weather and forecasting. 1997;12(4):736–741.

[38] Pinson P, McSharry P, Madsen H. Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography. 2010;136(646):77–90.

[39] Pinson P, Kariniotakis G, Nielsen HA, Nielsen TS, Madsen H. Properties of quantile and interval forecasts of wind generation and their evaluation. In: Proceedings of the European Wind Energy Conference & Exhibition, Athens; 2006. p. 1–10.

[40] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association. 2007;102(477):359–378.

[41] Morales JM, Conejo AJ, Madsen H, Pinson P, Zugno M. In: Renewable Energy Sources—Modeling and Forecasting. Boston, MA: Springer US; 2014. p. 15–56.

[42] Stephenson DB, Coelho CA, Jolliffe IT. Two extra components in the Brier score decomposition. Weather and Forecasting. 2008;23(4):752–757.

[43] Hersbach H. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting. 2000;15(5):559–570.

[44] Brier GW. Verification of forecasts expressed in terms of probability. Monthly weather review. 1950;78(1):1–3.

[45] Sweeney TAADR Dennis J ; Williams. Statistics. Encyclopedia Britannica; 2020 (accessed December 10th, 2020). https://www.britannica.com/science/statistics.

[46] Bruvik K. Power System Analyst, Statnett SF; December 4th 2020. Personal communication.

[47] Bentzien S, Friederichs P. Decomposition and graphical portrayal of the quantile score. Quarterly Journal of the Royal Meteorological Society. 2014;140(683):1924–1934.

A Appendix A: CRPS code

Listing 1: Functions used compute CRPS per instance.

```
def H(thresholds, actual):
    """Heaviside function. Makes CDF for observed value given:
    1D array of thresholds.
    1 float. Observed value.
    """
    result = [1 if t >= actual else 0 for t in thresholds]
    return result

def is_cdf_valid(case):
    """Checks if all probabilities are in within the interval [0,1]
       and CDF increasing, given:
       1D array of cumulative predictions, P(y <= t), for each threshold.
    """
    if case[0] < 0 or case[0] > 1:
        return False
    for i in range(1, len(case)):
        if case[i] > 1 or case[i] < case[i-1]:
            return False
    return True

def calc_crps(thresholds, pred, actual, width=1):
    """ Calculates the Continuous Ranked Probability Score given:
            1D array of thresholds.
            1D array consisting of rows of cumulative predictions,
                F(y <= t), for each threshold.
            1 float. Observed value.
            Threshold width. Default value at 1."""
    num_thresh = len(thresholds)
    num_pred=len(pred)
    crps = 0
    if (num_pred == num_thresh) and is_cdf_valid(pred):
        cdf_actual = H(thresholds, actual)
        for fprob, oprob in zip(pred, obscdf):
            crps = crps + (width*(fprob - oprob)**2)
```

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway