Distribution Network Expansion Cost Allocation Using Stochastic Cooperative Game Theory

Wonseok Choi, *Student Member, IEEE,* Hyeonjin Kim, *Student Member, IEEE,* Duehee Lee, *Member, IEEE,*

***Abstract*—In this paper, we propose a cost allocating methodology for net work expansion in distribution where renewable energy has penetrated. Common network expansion cost, including general distribution, are allocated by mathematically calculating contribution of the line. In case of distribution, expected demand of each consumption zone was determined and contribution calculated. However, the higher penetration ratio of renewable energy in distribution, the higher volatility of demand. Therefore, the cost allocating of existing methods does not accept this volatility, so it could cause irrationality if applied to current network. In addition, the postage-stamp method adopted by most countries is calculated only through amount of flow during peak, so it is not fair to allocate costs without considering overall benefits of network expansion.Therefore, we intend to solve the cost allocation problem by using stochastic cooperative game theory (SCGT) to consider above problems. To apply SCGT, we first make a demand map according to confidence interval through quantile regression using Gibbs sampling and XGBoost. Quantile regression allows us to calculate the utility benefits that take charge of demand region by outputting discrete probabilistic demand distribution over time. Such benefit calculation result in discrete representation of stochastic benefit associated with time for each period, and by substituting this for SCGT, we can get the cos t allocating results by time. To produce weighted mean according to peak of previously obtained cost allocating result, we use fuzzy weighted average (FWA) theory, and we can finally get cost allocation ratio that each utility should pay for network expansion**

I. Introduction

INDEPENDENT system operators (ISOs) should predict electricity loads so that the total generation amount from resource entities (REs) meets the total load from load serving entities (LSEs) in an electricity market [?]. ISOs can control generation amounts by giving dispatch orders to REs, but ISOs have minimal financial control over the load. The load is inelastic to the electricity price, and ISOs cannot easily change the electricity usage in real time [?]. Therefore, for stable system operation, the power system balances through demand forecasting, and for this purpose, a more sophisticated forecasting model is needed [?].

Load forecasting can be classified into short-term, mid-term, and long-term forecasting. In short-term forecasting, the load is forecasted based on the numerical weather prediction (NWP) up to two weeks ahead until the NWP is available. In mid-term forecasting, the load is forecasted between two weeks and three years. Finally, long-term forecasting covers the load forecasting for three years or more. Normally in long-term forecasting, in order to reflect the effect of human activity on the load in the future, the effects of economic growth should be considered [?].

Among the three different types of load forecasting, mid-term probabilistic load forecasting is the most beneficial to power system planning. For example, the required amount of ancillary services should be estimated a month ahead based on load forecasts, so short-term forecasting is not a proper approach for forecasting load, and long-term forecasting is not as accurate as mid-term forecasting. Furthermore, mid-term load forecasting is required to prepare unit commitment and maintenance schedules a month ahead of time based on the load forecasts [?].

While mid-term forecasting has many benefits, mid-term load forecasting is the most complex. Ordinary load forecasting models use NWP which includes weather variables like temperature, humidity, data, and time [?] that affect humans' load usage. If NWP data exists, the load can be predicted using the relationship between the NWP and the load [?]. However, NWP can only produce effective results when used for forecasts of up to two weeks. So, it is difficult to use NWP data for mid-term load forecasting.

When NWP data is unavailable, time series models have been used to forecast loads recursively [?]. Another method involves estimating future NWP scenarios from the historical weather data, and using these scenarios to predict future load [?].

Recently, probabilistic load forecasting has been proposed to circumvent unavailable NPW [?]. In probabilistic load forecasting, future NWP scenarios are synthesized and converted to future load scenarios, and then, point forecasts can be made based on the mean or median of the marginal distribution of the future load scenarios.

There is much literature about mid-term probabilistic load forecasting. One way to synthesize NWP scenarios is to reuse the historical data. In [?], ten years of the historical temperature data was used to generate 10 different temperature scenarios, which were then used to predict 10 different load forecasts. However, only a limited number of scenarios can be obtained from historically measured data, and this will reduce forecasting accuracy. In [?], to increase the number of scenarios, forecasting errors were collected by comparing the predicted values with the actual values. The errors were modeled as a normal distribution, where an infinite number of scenarios can be sampled. Many participants at global energy forecasting competition (GEFCom) 2017 used an hourly sampling method to synthesize temperature scenarios [?].

However sampling from one-dimensional distribution cannot consider relative humidity (RH) which has a significant effect on load forecasting [?].

Also [?] showed that the hourly or daily ahead data for temperature can be used to improve the preciseness of load forecasting. But, hourly samples cannot consider the joint probability between samples of different times. Also using hourly independent samples, matching time ahead data might deteriorate the correlation between features.

In this study, we develop an advanced algorithm for mid-term probabilistic load forecasting. First, as mentioned above, many studies used a one-dimensional temperature scenario, whereas we sample the interrelated temperature and dew point scenarios from their multivariate Gaussian distribution (MGD). In this process we also consider the daily patterns and dependency between hourly weather data using joint distribution of hourly random variables.

Second, we propose a probabilistic graphical model (PGM) which can provide an intuitive graphical model of the joint probabilities when considering a problem. PGM can visualize dependency and relationships between random variables and simplify marginal probability.

Third, we use Gibbs sampling. In the case of MGD, general sampling method is difficult to converge for a specific purpose in a multi-dimension distribution, and it is not possible to consider the dependency between variables. Therefore, we use Gibbs sampling in this paper to solve above problems.

Fourth, we use the extreme gradient boosting model (XGBoost). XGBoost is the ensemble algorithm that uses a combination of several decision trees. XGBoost has both performance and resource efficiency, so it is one of the most popular forecasting algorithms. XGBoost is an improvement on the gradient boost machine (GBM), and it uses several low accuracy forecast models rather than a single high accuracy model like GBM.

Fifth, We use kernel density estimation (KDE) for post-process which can generate smooth non-parametric distribution for hourly load from load scenarios which can preserve each observation. Also by using KDE we can get a mathematical form of load PDF which can be utilized in many different ways for power system problems. In this process we set the bandwidth parameter by forecasting residuals in the cross validation process which can substitute residual simulation.

Finally, we present a method of cost-allocation among utilities using cooperative game theory (CGT) to show the utilization of mid-term forecasting in electricity system planning based on the provided demand forecasts.

II. Forecasting Environment and Data Description

In this study, the load is forecasted by following the forecasting environment of the GEFCom 2017. The distributions of month-ahead load from 10 zones should be forecasted at every hour for 31 days upto two month ahead. The distribution is represented by nine quantiles from 10% to 90% at every 10%.

Furthermore, from the GEFCom 2017, the historical load from eight zones and historical weather observations from eight weather stations are provided. These data were hourly measured for 14 years. Historical weather observation consists of temperature and dew point. It should be noted that all training data is measured data, so the NWP is not used.

A historical load was measured from eight zones in the New England independent system operator (ISO-NE). Eight zones are New Hampshire (NH), Vermont (VT), Rhode Island (RI), west central Massachusetts (WCMA), south-east MA (SEMA), north-east MA (NEMA), Connecticut (CT), and Maine(ME). In addition to eight zones in ISO-NE, the entire ISO-NE is considered as a hierarchical structure total load.

The training set consists of the weather and time data. As mentioned above the weather data includes temperature and dew point, and the time data includes year, month, day, and hour.

The 8 different forecasting models predict the load of 8 zones in parallel manner. Because no weather station is representative for ISO NE and total MA zone thus we added forecasted values underlying each hierarchical structures.

For training set and NWP scenarios, selection of weather station is determined by gamma test. Selected weather stations for each zone consist of joint MGD and Gibbs sampling through Probabilistic Graphical Model is utilized to generate target input stream which is NWP scenarios.

In this paper, the performance of distribution is evaluated by the pinball loss function. The pinball loss in [**?**] is used to measure the performance of probabilistic forecasting when the distribution is represented as quantiles. The loss function is defined as:

$$L(q_a, y) = \begin{cases} (1 - \frac{a}{100}) \times (q_a - y) & if\, y < q_a \\ \frac{a}{100} \times (q_a - y) & if\, y \geq q_a, \end{cases} \quad (1)$$

where a represents a quantile percentage, q_a represents the quantile prediction, and y represents the target value.

III. Scenario Generation

In this section graphical modeling and Gibbs sampling are introduced. Observed temperature and dew point are designed as a MGD based on the graphical modeling, and their scenarios are synthesized based on the Gibbs sampling.

A. Probabilistic Graphical Modeling

The graphical model represent the multivariate probabilistic distribution as a network graph of nodes and edges. The nodes represent variables in the multivariate distribution, and the edges represent the dependency between nodes. The direction of an arrow in an edge represents a dependency. Suppose there are two variables A and B. If an edge has an arrow from A to B, B depends on A. By following arrows in a network, we can build the joint distribution. For example, starting from a root A with a priori probability $p(A)$, we can build the conditional probability of B with respect to A as $p(B|A)$ by following an arrow. Then, we can build the joint distribution $p(A, B)$ over the variables A and B as the product of $p(B|A)$ and $p(A)$. In other words, in a network, a multivariate distribution can be decomposed by a multiplication of a few conditional probabilities by following arrows. Finally, the probabilistic model in a network helps us to sample scenarios.

The graphical model provides Three strong advantages to handle multivariate distributions. First, it provides a simple way to visualize the structure of a probabilistic model since

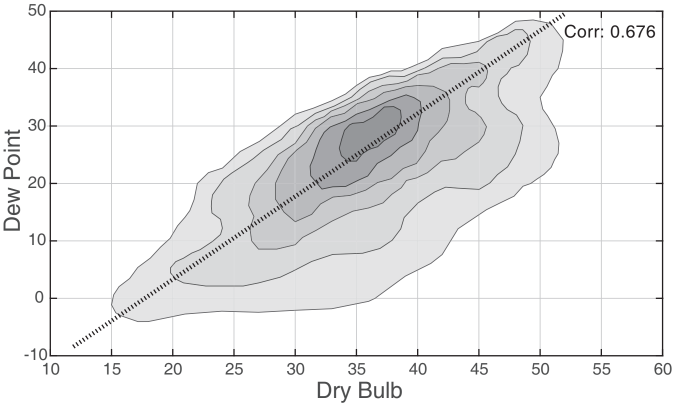

Fig. 1. Non-parametric bivariate distribution of dew point and dry bulb is represented. Two random variable show high correlation, which makes weather scenario generation from joint density of these variable more convincing.

it is easy to detect the conditional independency through the network. For example, when a series of events occurs in sequence, the graphical model can represent a probability of the last event with respect to the probabilities of only dependent events. Second, since a network represents only the dependent variables, less number of parameters is required than a full description of distribution. On the contrary, without the graphical model, the probability of the last event should be represented by the probabilities of all beforehand events. Third, we can utilize the conditional independency. For example, in our application, if we know the date, the temperature from two different zones on the same date are independent. It is called a conditional independency. Since there is no strong reason that the temperature from different zones depend each other, we can assume that temperature is independent on the condition of the given day. By utilizing this conditional independency, we can simply multiply the probability of temperature and dew point.

In our application, we use the graphical model to design the multivariate distribution of load and training data and sample future scenarios of temperature and dew point from the multivariate distribution. For our application, the temperature and dew point temperature depends on a month of year, and temperature and dew point have dependency on each other on given month. With a pair of temperature and dew point temperature we can derive a relative humidity which is not observed. Based on the reasonable inference, we can deduce a network graph from a given multivariate distribution. A multivariate distribution using graphical model that we considered is plotted in Fig 1. With graphical model we sampled multiple temperature and dewpoint scenarios depend on given month and time. Using two variables, temperature and dew point, we formed the multi probabilistic distribution.

B. Historical data collection for scenario generation

The probabilistic model can be estimated from observation. We have temperature and humidity data from 8 different zones for 14 years. However, at the given day of year and hour, there are not many samples to build the probabilistic model. Besides, for the multivariate distribution, more samples are required to describe the distribution.

Therefore, we collect data for a few hours ahead and a few hours later. At the same time, we collect data for a few days ago and a few days later. The way to estimate accurate time and day spans were studied in [**?**]. We collected dewpoint and drybulb three hours ahead and after from each 14 years of historical data. A similar study for our application should be performed in future work.

C. Gibbs Sampling

In this study, the Gibbs sampling is used to sample sets of temperature and dew point for each zone. The Gibbs sampling is a kind of Markov chain Monte Carlo (MCMC) sampling algorithm, and it can also be seen as a special case of the Metropolis-Hastings sampling algorithm [**?**] . As well as the Gibbs sampling, there are several sampling techniques in a graphical modeling: ancestral sampling, importance sampling, likelihood-weighted Sampling, logic sampling, Rao-Black wellisation, and Gibbs sampling. Among these techniques, the Gibbs sampling is used in our application since it is simple and fast [**?**].

The main idea of the Gibbs sampling is as follows. Since it is difficult to sample a set of variables from a multivariate distribution, we can sample each target variable from the conditional distribution of a target variable with holding the others constant. After a variable is sampled, we will sample the next variable based on other sampled variables. In other words, we sample $x_1^{(j+1)}$ from $\pi(x_1^{(j+1)}|x_2^{(j)}, x_3^{(j)}, \ldots, x_d^{(j)})$. Then, we sample $x_2^{(j+1)}$ from $\pi(x_2^{(j+1)}|x_1^{(j+1)}, x_3^{(j)}, \ldots, x_d^{(j)})$. It should be noted the index of each variable. Finally, if we continuously sample variables one by one, the set of samples will follow the joint distribution. In this process, we assume that the dependence of samples follows a Markov chain, since the Gibbs sampling is a kind of MCMC sampling algorithm.

The advantage of using the Gibbs sampling shed when the high-dimensional full joint distribution is not represented analytically, but the conditional distribution is simply described. Besides, if other variables are independent to a target variable, we can simply the conditional distribution, and it becomes a low-dimensional conditional distribution with only related variables.

There are many variables from 10 zones, but we can reduce the number of variables in the conditional distribution through our graphical model. Therefore, we only need to build the Markov chain for each zone. Therefore, we secure the conditional independency between scenarios from different zones through the graphical model. The temperature and dew point scenarios in different zones are sampled individually under the condition of target time. For each zone, scenarios are synthesized together through the Gibbs sampling.

The algorithm of the Gibbs sampling is described below. In this study, we want to sample from a joint distribution $P(\mathbf{x}) = P(x_1, x_2)$ where x_1 represents the temperature, and x_2 represents the dew point. We assume that $P(\mathbf{x})$ follows the bivariate Gaussian distribution. First, we initialize samples.

$$\mathbf{x}^0 = (x_1^0, x_2^0) \tag{2}$$

```
Data: Temperature and dew point
Result: $N$ sets of samples
for i:1 to $N$ do
    Iinitialization $\mathbf{x}^0 = (x_1^0, x_2^0)$
    for j:0 to $M$ do
        Sample;
        $x_1^{(j+1)} \sim \pi(x_1^{(j+1)}|x_2^{(j)})$ ;
        $x_2^{(j+1)} \sim \pi(x_2^{(j+1)}|x_1^{(j+1)})$ ;
        Go to the next iteration;
    end
    Collect $(x_1^M, x_2^M)$ ;
    Go to the next iteration;
end
```

Algorithm 1: The algorithm of the Gibbs Sampling

Then, we sample x_1 while holing x_2 constant as an initial value.

$$x_1^{(1)} \sim \pi(x_1^{(1)}|x_2^{(0)}) \tag{3}$$

Then, while holding x_1 as a sampled value, we sample x_2.

$$x_2^{(1)} \sim \pi(x_2^{(1)}|x_1^{(1)}) \tag{4}$$

It should be noted that this procedure can be generalized for arbitrarily many variables. In this study, we assume that the temperature and dew point follow the bi-variate Gaussian distribution. The conditional Gaussian distribution when the other variable is known is introduced as

$$\pi(Y|X=x) \sim N\left(\mu_Y + \rho\left(\frac{\sigma_Y}{\sigma_X}\right)(x-\mu_X), \sigma_Y^2(1-\rho^2)\right). \tag{5}$$

The joint Gaussian distribution is estimated from the observation of temperature and dew point. These processes should be repeated for several times until the conditional dependency between variables are represented in samples, and the independency of initial values is disappeared. The processes can be summarized as an algorithm in the Algorithm 1, where N is the required number of samples, and M is the number of iterations.

In this application, we assume that vectors of samples follow the multivariate Gaussian distribution.

D. Scenario Generation

In this section, we will generate temperature and humidity scenarios for target month. We considered multivariate joint distribution which is consisted of random variables of few different weather station's dry bulb and dew point temperature. Each dew point and dry bulb temperature is a vector of random variables which consist of 24 random variables that mean a variable for an each hour. Sampling NWP from this joint distribution can consider correlation between dry bulb and dew point temperature and this pair of data can be interpreted as relative humidity. The advantage our relative humidity on load forecasting is described in []. Temperature and humidity is continous in every moment thus sampling for each hour and combining them arbitrarily can break the pattern which has in nature. Through our method such pattern can be kept. This is important because considering preditors such as temperature of few hour ahead can improve forecasting performace which is written in[]. And if temperature scenarios which don't consider the correlation of each hour make awkward match in predictor set. For checking our sampling method, we considered sampling from other distribution. We made each distribution of DB and DP temperature considering hourly variable and randomly matched each sample and we call this as individual method. We also considered sampling from hourly distribution of DP and DB and this is hourly method.

IV. Forecasting Methods

In this section, the distributions of load forecasts are estimated for each scenario through two different methods, which are the qunatile regression of gradient boosting machine and the quantile regression of random forest. Then, the multiple distributions of scenarios for a single target case are combined through the kernel density estimation based on the Gaussian mixture model. Furthermore, The performance of the Gaussian mixture model is compared to the empirical method of distribution averaging.

A. Basic of Quantile Linear Regression

In this subsection, we will introduce how to perform the quantile regression.

The quantile y_q at the given percentage q means the value of random variable whose cumulative probabilities become the given percentage. This can be represented as

$$y_q = F^{-1}(q) \tag{6}$$

where q represents the percentage, and F represents the cumulative distribution function (CDF). For example, the median is the 0.5 or 50% quantile. Therefore, the quantile is the output of inverse CDF at the given percentage.

We can start from a well known fact that the value, which can minimize the sum of absolute errors between samples, is the median. In short, the median has the least-absolute deviation (LAD). It is obvious that finding the median is a sorting and selecting problem. Since $P(X \leq m) = P(X \geq m) = 0.5$, where m is the median, the median is in the middle of sorted samples. Then, if we assume that we can only find a median, in order to find other quantile for different percentages, we can change the number of samples in the left hand side of quantile by multiplying a weight factor. For example, if we want to find the quantile at the 80 %, we can multiply the coefficient 0.8 to the right hand side of quantile. This will make a virtual situation that we have many samples right hand side of quantile. Therefore, finding a median is to find the value of the LAD. Similarly, finding a quantile is to find the value of the LAD of weighted samples.

By using this approach, we can find a quantile at a given percentage, which can minimize the sum of LADs. Generally, the ordinary linear regression minimizes the sum of all losses, but the quantile regression minimizes the asymmetric penalties with different weight factors. When it is over-predicted, the $1-q$ is multiplied to the loss, and when it is under-predicted, the q is multiplied to the loss. Then, the loss function becomes as

$$\begin{aligned} L(y, \beta_q) = q \sum\nolimits_{y_i > \mathbf{x}_i \beta_q} (y_i - \mathbf{x}_i \beta_q) \\ + (1-q) \sum\nolimits_{y_i \leq \mathbf{x}_i \beta_q} (\mathbf{x}_i \beta_q - y_i) \end{aligned} \tag{7}$$

where β is the coefficient of linear regression.

The quantile regression represents the distribution of forecast as quantiles.

In the linear regression, it leads to the projection of solution vector onto the subspaces of observation data. On the contrary, in the quantile regression, it should be solved by techniques in the linear programming.

If we know the CDF of a certain variable, we can get quantiles by inversing the CDF. However, if we only know samples, we can get quantiles by solving an optimization problem for the given samples. Moreover, the quantile regression can be considered as inversing the CDF if we can get the empirical CDF from samples.

B. Quantile Regression of Gradient Boosting Machine

The GBM is the most widely known a probabilistic regression tool [**?**]. Several forecasting competition winners have been used this algorithm.

There are two representative methods to forecast the distribution of forecasts: the quantile regression based on a specific forecasting algorithm and the estimation of distribution through analyzing forecasting errors. The second method estimates the error distribution by collecting forecasting errors. Errors are collected by comparing the actual values and forecasted values.

However, the quantile regression can generate different distributions for the given weather conditions. In other words, the error distribution methods can have limited number of different distributions for different given observation data. On the contrary, quantile regression can have different variance for the given observation data. In other words, if the quality of the target observation data is bad, the variance of the distribution increases, but if the quality of the target observation data is good, the variance of the distribution decreases, and it means that we have a narrow and accurate distribution.

For the quantile regression approaches, the representative forecasting algorithms used in the quantile regression approaches are the GBM and the RF. In this study, we use both methods and compare their performance.

The problem of the quantile regression is that we should solve the same number of problems for the given number of quantiles. Furthermore, quantiles might not be sorted, so the CDF might not be non-increasing, then the PDF could be negative, which cannot happen normally.

In this study, we define the "training set" as a combination of "N training data" $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ and the "N target values" $\mathbf{y} = \{\mathbf{y_1}, \ldots, \mathbf{y_N}\}$. The goal of the forecasting is to estimate the function $\hat{F}(\mathbf{X})$, which can approximate the actual function $F(\mathbf{X})$ so that $\mathbf{y} = F(\mathbf{X})$. The $\hat{F}(\mathbf{X})$ can be estimated to minimize the sum of loss function $L(\mathbf{y}, F(\mathbf{X}))$ over the combination of $(\mathbf{y}, \mathbf{X})$. Then, the $\hat{F}(\mathbf{X})$ can be found by solving a simple minimization problem

$$\hat{F}(\mathbf{X}) = \arg\min_F \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)). \tag{8}$$

In the GBM, this minimization problem is solved by three key ideas. The first key idea of the GBM is that we can present the $\hat{F}(\mathbf{X})$ as the sum of several functions $f_m(\mathbf{X})$ as

$$\hat{F}(\mathbf{X}) = \sum_{m=1}^{M} f_m(\mathbf{X}), \tag{9}$$

where $f_m(\mathbf{X}) = b_m h(\mathbf{X}; \mathbf{a}_m)$ under the assumption that all $f_m(\mathbf{X})$ have the same basis function $h(\mathbf{X})$ [**?**].

Then, the second key idea, which is a strategy to find $\hat{F}(\mathbf{X})$, is to update it by finding $b_m h(\mathbf{X}; \mathbf{a}_m)$ as

$$F_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) + b_m h(\mathbf{X}; \mathbf{a}_m) \tag{10}$$

where $h(\mathbf{X}; \mathbf{a})$, which is a basis function with known parameters, is called a weak learner. The $(b_m, \mathbf{a}_m)$ can be obtained by solving

$$(b_m, \mathbf{a}_m) = \arg\min_{b, \mathbf{a}} \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + b h(\mathbf{x}_i; \mathbf{a})). \tag{11}$$

The updating component $b_m h(\mathbf{X}; \mathbf{a}_m)$ in (10) can be found through the steepest descent approach [**?**], where $h(\mathbf{X}; \mathbf{a}_m)$ is defined as the gradient of the loss function as

$$g_m(\mathbf{x}_i) = \left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]. \tag{12}$$

However, the problem of this approach is that the gradient is obtained only for the observation data $\mathbf{x}$. The solution of this problem is to find the smooth functoin $h(\mathbf{X}; \mathbf{a}_m)$ so that we have values for unknown observation $\mathbf{x}'$. One of good smooth basis function is a decision tree [**?**], and we will also use a decision tree in this study. Then, what we have to do is to find parameters $\mathbf{a}_m$ so that $h(\mathbf{x_i}; \mathbf{a}_m)$ has a similar values to $g_m(\mathbf{x}_i)$. However, since it is difficult to find b_m and $\mathbf{a}_m$ simultaneously, we will find $\mathbf{a}_m$ first for the given b_m and find b_m. The $\mathbf{a}_m$ can be found by solving the next minimization problem

$$\mathbf{a}_m = \arg\min_{b, \mathbf{a}} \sum_{i=1}^{N} \left[-g_m(\mathbf{x}_i) - b h(\mathbf{x}_i; \mathbf{a}) \right]^2 \tag{13}$$

This is a simple regression tree generation. Then,

$$b_m = \arg\min_{b} \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + b h(\mathbf{x}_i; \mathbf{a}_m)) \tag{14}$$

Finally, we could estimate a set of $(b_m, \mathbf{a}_m)$.

These all processes depend on the loss function. For example, if the loss function is a simple least-square loss function, which is defined as

$$L(y, F) = \frac{(y - F)^2}{2}, \tag{15}$$

then the gradient can be simply calculated as

$$g_m(\mathbf{x}_i) = y_i - F_{m-1}(\mathbf{x}_i) \tag{16}$$

In this study, we should use the quantile loss function, which is defined in [?] as

$$\begin{aligned} L(y, F) = \alpha \sum\nolimits_{y_i > f(\mathbf{x}_i)} (y_i - F(\mathbf{x}_i)) \\ + (1 - 1\alpha) \sum\nolimits_{y_i < f(\mathbf{x}_i)} (F(\mathbf{x}_i) - y_i) \end{aligned} \tag{17}$$

where α represents the quantile range, i.e. 0.1 represents 10%. Then, the gradient can be obtained as

$$g_m(\mathbf{x}_i) = \alpha \mathbf{I}(y_i > f(\mathbf{x}_i)) - (1 - \alpha)\mathbf{I}(y_i \leq f(\mathbf{x}_i)) \tag{18}$$

Finally, when the GBM forecasts the future value for new data, GBM generates outputs by adding results of all basis functions.

C. Kernel Density Estimation

In this subsection, we will find the final list of quantiles by combining lists of quantiles from scenarios. We use the kernel density estimation (KDE) with a Gaussian kernel. The Gaussian mixture model is a popular way to estimate the multi-modal distributions, which has multiple peaks when we have several distributions of different shapes [?]. Besides, by building a final list of quantile, we can have a non-decreasing list of quantiles. We also build the final list of quantiles by simply collecting samples from each list of quantiles.

The Gaussian mixture model (GMM) is another way to combine several distributions. It is similar to the KDE, where all points have a same small distribution. On the contrary, in the GMM, several Gaussian distributions with different means and variances are fitted to samples, and the number of distributions should also be estimated. In this application, since each quantile does not follow the Gaussian distribution, the GMM might not fit to our application.

The basic idea of the KDE in [?] as follows. In the KDE, the probability density of a new random variable x is measured by adding all numbers of samples within the fixed range around x. The number of samples K can be calculated as

$$K = \sum_{n=1}^{N} \kappa \left(\frac{x - x_n}{h} \right), \tag{19}$$

where N is the number samples, x_n is the nth sample, κ is the kernel function of counting, and h is the smoothing parameter. The h trades off between the variance and bias [?]. A large h leads to a very smooth distribution, but a small h leads to a noisy distribution, where each variable has a peak in the worst case. It should be noted that x is a scalar variable. By normalizing it, we can get the probability density of x as

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} \kappa \left(\frac{x - x_n}{h} \right), \tag{20}$$

where D is the dimension of data, so it is one in this application. When the kernel function is a Gaussian distribution, the marginal probability over samples is defined as

$$P(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\sqrt{2\pi h^2}} \exp \left\{ - \frac{\|x - x_n\|^2}{2h^2} \right\} \tag{21}$$

Therefore, the final list of quantiles is obtained by placing a Gaussian over each sample and adding distributions over all samples. The sum of distributions should be normalized by dividing the number of all samples.

In order to test the performance of the KDE, we just simply generate the final density of forecasts by adding a single quantile from each scenario. The final quantile can be generated from a quantile of each scenario. For the given adjacent two quantiles among nine quantiles from 0.1 to 0.9, we can generate uniformly distributed samples. Then, we can have samples from the given list of quantiles. If we collect all samples from all lists of quantiles. For the collected samples, we can re-build a list of quantiles from 0.1 to 0.9.

V. Cost Allocation

With the arrival of the modern electricity market, the cost allocating method of public-use facilities such as power transmission lines are becoming more complicated [?]. For example, in the past, the cost allocation of power transmission lines were carried out by calculating the amount of flow at each plant based on the peak time [?]. However, revenue from the power system not only arises from increased flow, but also directly or incidentally, such as increased system stability and flexibility. This approach is inefficient and incurring unnecessary costs, although it may be the fairest way to examine all structures that generate revenue in many ways and assess contributions. Thus, today's cost allocation approach have become more complex and equitable, such as a calculating network usage using impedance [?], an AC power flow Sensitivities [?], and a game theory [?].

A. Cooperative Game Theory

In this paper, we use *cooperative game theory* as a cost allocation method based on benefit. Generally, game theory is divided into two. One is *cooperative game theory* (CGT) and the other is *non-cooperative game theory* (NCGT). NCGT do not seek collective gain through discussion or cooperation between the parties, but consider probabilistic value according to personal choice. The CGT can discussion each other. Thus, interested parties form aggregation called *coalition*, pursue common benefit. Because of characteristic that participants can discussion, it can deduce more accurate result [?]. The possibility of discussion should be carried out with the game theory considering the combination of participants. Progressing the game theory in all cases can reveal the contribution of each participant in the game, enabling intrinsic characteristics. The structural features of the CGT used in the paper are as follows:

1) Core: Core which is the concept expanded *Pareto efficiency* includes feasible set by some axioms.

$$\sum_{i \in N} x_i = v(N) \tag{22}$$

N is the universal set that includes all participants. $v(N)$ is the coalition function, represent the worth set has. The first means that all benefit is allocated to all participants. And this equation is divided two inequalities.

$$\sum_{i\in N} x_i \geq v(N) \tag{23}$$

Second equation means feasibility. The participants can't allocate more than they produce.

$$\sum_{i\in S} x_i \geq v(S) \tag{24}$$

where *S* is the subset of *N* and means coalition. The last axiom is that the participants in coalition should get pay-off at least they produce. The *Core* solves the problem using optimization about this axioms. However, if you proceed optimization for the *Core*, it can appear several value. To allocate clearly, it needs the unique. Thus, the *Nucleolus* is introduced.

2) Nucleolus: Nucleolus is the concept expanded at the *Core*. In *Nucleolus*, the focus of the optimization is set to minimizing the worst inequity. To satisfy this, the *excess* is adopted, *excess* represent the difference of worth and sum of pay-off in coalition:

$$e(x, S) = v(S) - \sum_{i\in S} x_i \tag{25}$$

3) Optimization: When the worth function was set, proceed the optimization . The optimization proceed is as follows.

$$\begin{aligned} &\text{Max } \varepsilon \\ &\quad x_i \geq B(i) + \varepsilon \\ &X_i(s) \geq B_i(s) + \varepsilon \\ &\sum_{i\in S} x_i = B(N) \end{aligned}$$

Except for *excess*, all of them is constraints for *Core* axioms. Calculate benefits before and after forming coalition and set constraints according to axioms. Constraints for personal just set condition as many as the number of participants, but in coalition should calculate all possible parties. If there are *N* participants in CGT, it set 2^n-2 coalitions that except when all participants are included and excluded from 2^n that can combine all cases. Once proceed optimization, the results come out. Commonly the results may be not unique. In this happen to get unique solution, proceed optimization again. At preceding optimization, decide minimum value for x_k(= $\hat{x}_k$) and ε(=ε_k), apply to the equation as follows.

$$\begin{aligned} &\text{Max } \varepsilon_2 \\ &\quad x_k = \hat{x}_k \\ &\quad x_i \geq B(i) + \varepsilon_1 + \varepsilon_2 \\ &X_i(s) \geq B_i(s) + \varepsilon_1 + \varepsilon_2 \\ &\sum_{i\in S} x_i = B(N) \end{aligned}$$

Repeat this procedure maximum *N*-1 until get unique.

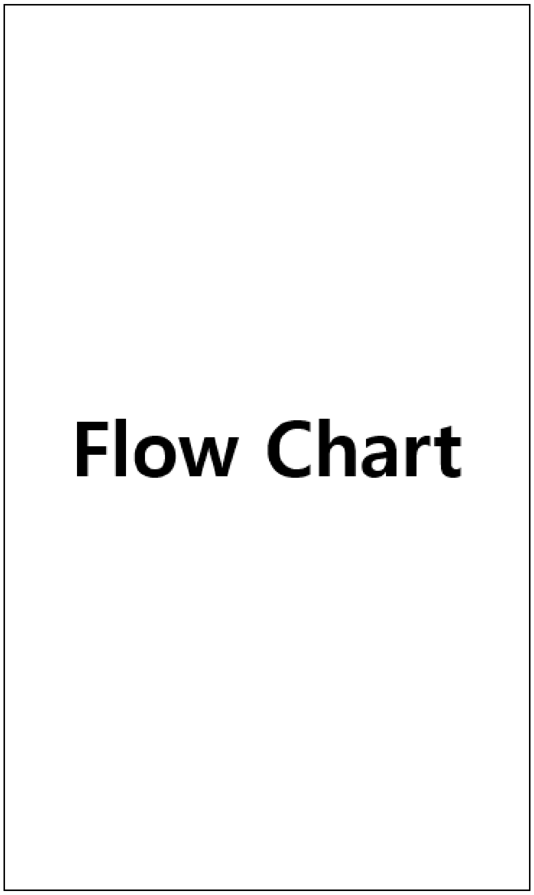

Fig. 2. flow chart

B. Cost Allocation Methodology

The cost allocation methodology proposed in this paper consists of a series of steps. The first step is transmission expansion planning (TEP). The new transmission lines to be built will benefit economically only if the benefits from long-term operations are greater than the cost of construction [**?**]. Therefore, to see these economic effects, additional power transmission facilities need to determine the buses and the power transmission capacity that can reduce the incidental costs considered reserve, outage and congestion rents incurred by the system as a whole. Based on this, the following optimization expression is established:

$$arg \min_{\zeta(r,s,t)} \sum_{r,s,\in K} (\pi_\zeta + \psi_\zeta) \tag{26}$$

Where *K* means set of all nodes in system, *r,s* are selected nodes, *t* is line capacity and ζ means order pair. π of the objective function means the costs incurred when constructing transmission lines on each established node and ψ means the incidental costs incurred in the same situation.

After TEP, the transmission line's node and capacity were determined, the benefit generated between each participant should be calculated. In system congestion situations, system operators direct generators to reduce generation and replace them with regional development for system reliability and stability. Regional generation creates relatively high prices, resulting in an increase in local marginal price (LMP) and overall unnecessary maintenance costs. In this situation, if new power transmission line is installed, the system stability will be increased again and additional power generation of existing power plants will be possible. At the same time, LMP is lowered and maintenance costs are reduced.

Specifically, power plant (PP) benefit from additional generation, transmission system operator (TSO) benefit from more transmission charges and reduced system reliability costs. LSE will able to use electricity that is as cheap as LMP's profit. However, although the TSO and LSE can calculate direct benefit through profit, there is an unreasonable aspect to simply calculate benefit through power generation growth for PP.

Even if the generation has increased, it is not possible to require high cost allocating to the plant if the utilization rate of the power transmission line is low. to compensate for this, the plant calculates the gain of the increase in power generation over the newly constructed lines by multiplying the increase in power generation power transfer distribution factors (PTDF).

Finally, based on the benefits calculated in the previous step, proceed a calculation considering all coalition. $\boldsymbol{S}(\mathbf{S} \in \mathbf{R}^n)$ is the universal set which include all participants that have revenue structure. The first described CGT optimization expression is represented by a matrix form:

$$\begin{aligned} &\max \varepsilon \\ &\quad \mathbf{X} \geq \mathbf{\Lambda} \cdot \mathbf{B} + \varepsilon \\ &\sum_{i \in S} x_i = B(S) \end{aligned} \tag{27}$$

where $\mathbf{X}$ has $2^n \times 1$ dimension, means the value allocated according to each coalition. $\mathbf{\Lambda}$ and $\mathbf{B}$ mean whether to participate and benefit each element respectively. Solving the given optimization results in a allocate value for each participant, and the final cost allocating is derived by converting the share value into a percentage and then multiplying the cost to be allocated.

$$\nu(x_i) = \frac{x_i}{\sum x} \times TC \tag{28}$$

ν means allocated cost and TC means total cost that should be allocated.

VI. Simulation

In this section, we proceed numerical experiments on a mid-term forecasting and cost allocation assuming a transmission system for specific zone.

A. Forecasting Architecture

For variable selection, we used to build proper regression model, we first tested point forecasting performance. We compared the performance by two different direction. First by cross-validation performance of given training data we tried to find best parameter of regression model. Also in order to confine the scope of usage of given observations, performance of given training data was constructed by iteration of adding data set. Being based One year ahead observations from the target month and a year when the average demand and temperature were similar to right before the target month, training data were constructed by adding some yearly data. After building learning model, we fed test data altering weather data for 1000 thousand number of scenario. Then we have 1000 load forecasting observations and 9000 observations for quantile regression each hour. To present load distribution KDE method is used.

TABLE I
Forecasting Model Comparisons

Zone number	Zone Name	Our Performance	Benchmark
1	CT	97.79	98.8
2	ME	23.31	23.88
3	RI	19.17	21.53
4	NH	29.14	29.64
5	WCMASS	49.5	55.25
6	SEMASS	44.88	49.51
7	NEMASS	70.93	73.16
8	MASS	161.11	175.86
	ISO NE	330.4	351.70

In each process of generating scenario and distribution fitting we set different methods to compare the performance. For sampling method we made distribution of DB and DP independently, considering hourly variable and randomly paired each sample and we call this as individual method. We also sampled from hourly joint distribution of DP DB and this is hourly method. Also to compare the method of choosing final load distribution of each hour, we compared KDE to picking up each 9 quantiles from empirical distribution of 1000 observations.

B. Performance of Hierarchical Forecasting

We measure the performance of hierarchical forecasting. In our hierarchical forecasting, we forecast the load of area whose forecasting performance is better than other areas. However, the performance of hierarchical forecasting was worse than individual forecasting.

The second tested technique is used to forecast the total MA and NE-ISO. First, we forecast the load of all individual areas and add them to forecast the NE-ISO. Second, we directly forecast the load of MA and NE-ISO. Finally, the performance of individually forecasting is better than a direct forecasting of a total value.

C. Performance of Forecasting

We compared performance of 10 different forecasting model by pinball loss function. Each score is a summation of forecasting loss of 8 different zone, MA and whole ISO NE. As in Table II model based on testing scenario which considered dependency of two different temperature and hourly correlation and XGboost with KDE showed best performance. GBM quantile regression showed in Fig 3.

D. Cost Allocation

We supposed 8-nodes system which has 8 branches and base plants were located in 1, 2, 3 node respectively (called GEN1, GEN2, GEN3). In addition, each node has a sub generator to

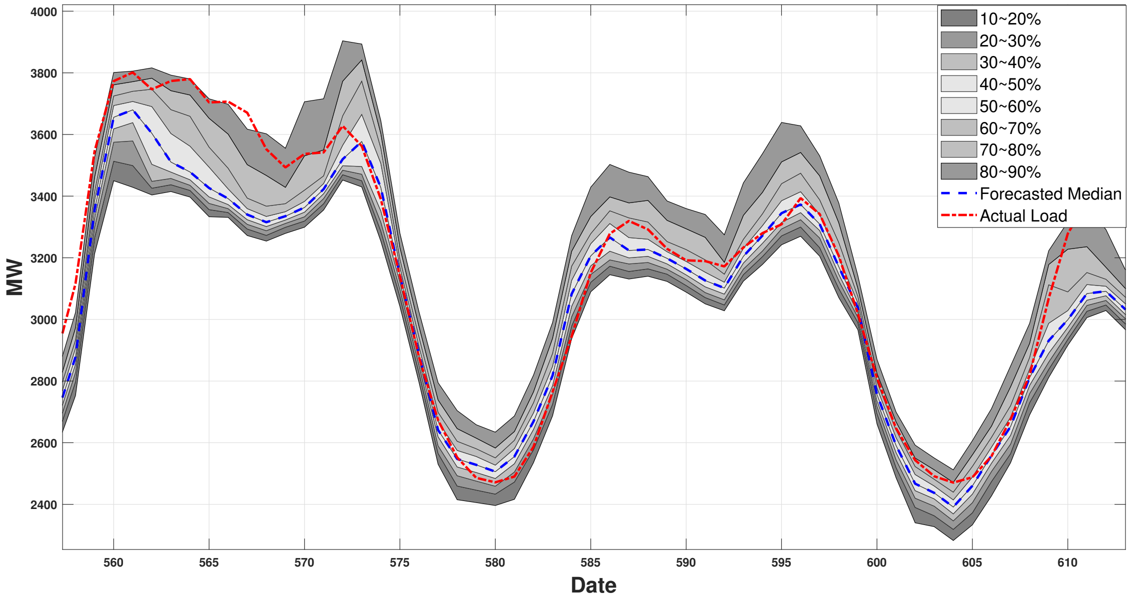

Fig. 3. Gaussian Mixture figures with overlapped distributions

TABLE II
FORECASTING ACCURACY, PROFITABILITY AND COMPUTATIONAL TIME OF THREE PROBABILISTIC SOLAR POWER FORECASTING MODELS

	Proposed	**Benchmarks**	
	Two-Stage Prob.	PP	QGB
Pinball Loss	**0.245**	0.300	0.258
Entropy	3.652	4.035	**3.612**
DA Revenue (KRW)	2,071,462	1,564,340	2,321,903
RT Penalty (KRW)	417,755	772,181	743,511
Profit (KRW)	**1,653,707**	792,158	1,578,392
Computation time (min)	58.75	**0.01**	60.53

prevent imbalance, and the sub generator has a higher power generation cost than the base generations. Fig 4 shows a entire system and, Table III and Table IV denote detailed system parameters.

TABLE III
BRANCH INFORMATION

Bus		Rating(MW)	Reactance(Ω)
to	from		
1	4	2000	0.0576
4	5	1500	0.0920
5	6	1500	0.1700
3	6	2000	0.0586
6	7	1500	1.008
7	8	3000	0.0720
8	2	1500	0.0625
2	4	2000	0.0780

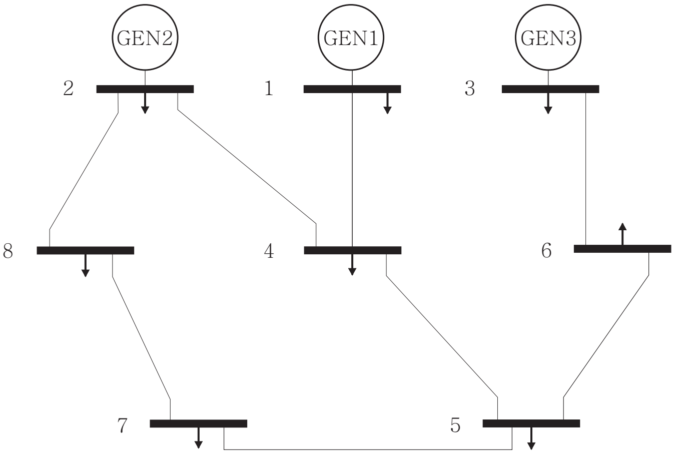

Fig. 4. Eight-bus system which describes base generators except sub generators

TABLE IV
GENERATOR INFORMATION

Gen type	Bus	Labeling	Capacity(MW)	Cost coefficient		
				a	b	c
Base Gen	1	GEN1	5000	0.1100	5.00	150
	2	GEN2	8000	0.0850	1.20	600
	3	GEN3	6000	0.1225	1.00	335
Sub Gen	1	SUB1	2000	1.7750	10	150
	2	SUB2	2000	2.2123	10	150
	3	SUB3	2000	1.8548	27	150
	4	SUB4	2000	2.3924	50	150
	5	SUB5	2000	1.9680	46	100
	6	SUB6	2000	1.7879	48	100
	7	SUB7	2000	1.8779	10	100
	8	SUB8	2000	2.0451	60	100

In generator, the GEN2 has cheapest unit price, it maintain a similar generator to others. Because 2-8 transmission line has congestion, GEN2 directly connected to this have no choice but to place a limit on the generating. Simultaneously the LSE use the electricity at higher price, if this status is maintain, losses of total system accumulate.

To decrease system losses, it need additional transmission line. Once proceed TEP optimization, the added transmission line constructed in 2-8 branch. 2-8 branch decrease the LMP and solve existing congestion. Thus we assume this content and keep experiment.

After construction, we calculate the power variance using OPF. According to OPF, the power of GEN2 larger. In contrast the GEN1 and GEN3 decrease for static demand. In this paper, we consider only positive effect of addition line neglect under the zero. The table below is the detailed lineage contents used for OPF.

To decrease system losses, it need additional transmission line. Once proceed TEP optimization, the added transmission line constructed in 2-8 branch. 2-8 branch decrease the LMP and solve existing congestion. Thus we assume this content and keep experiment.

After construction, the OPF for one month is calculated to confirm the *flow* on the expanded transmission line, and the contribution of each generator to the line is calculated by multiplying the PTDF. The benefit of the final generator is calculated by substituting the *flow* of the lines by each generator into the cost curve.

Now that we have obtained the generator benefit, we calculate the benefit of LSE and TSO. The LSE can be calculated through the Lagrangian multiplier of OPF, and the upper limit (850$/MW) of the LMP was set prior to line expansion to avoid the dominant responsibility of the LSE dut to the LMP decline. The benefit of TSO can be calculated by the wheeling rate(10$/MW).

Next, the CGT is applied based on the calculated benefits. Five institutions contributed to the cost of expansion transmission lines, with five participants in game theory, and an optimization for this is established.

$$
\begin{aligned}
&\max \varepsilon \\
&\text{subject to.} \sum_{\mathbf{I}\in S} \mathbf{X}_I \geq \sum_{\mathbf{I}\in S} \mathbf{B}_I + \varepsilon \\
&x_i \geq B_i + \varepsilon \\
&\sum_{i=1}^{5} x_i = \sum_{i=1}^{5} B_i
\end{aligned}
\tag{29}
$$

where $S(S \in R \rightarrow R^{2^5-7})$ means all coalitions and **I** is possible set in S. **X** is subset of all the coalition and is the number of possible vectors. The constraints were expressed based on the CGT *core* described above.

The results of game theory on the monthly average benefit are as follows:

TABLE V
COST ALLOCATION RESULTS

GEN1	GEN2	GEN3	LSE	TSO
0.2774	42.0559	0	49.3645	8.3022

Thus, Multiplying the ratio shown in the table by the cost of expanding the transmission line finally determines the cost to be shared by each institution.

VII. Conclusion

In this paper we proposed novel method to generate NWP scenario for Mid-Term probabilisic load forecasting. We compared the forecasting performance between each scenario making methods. Scenario considering a distribution of 24 hours random variable of each dew point and dry bulb temperature outperformed other methods such as individual sampling or hourly sampling and competition benchmark. We used XGboost for regression model which is most popular machine learning method in recent years for powerful regression ability and low computational load. We also described a method of generating final load distribution from observations of thousand differnet scenarios using kernel density estimation. By this we could get a mathematical form of distribution and also keep original distribution considering each observation.