**Responses to Referees' Comments on Paper TSG-00141-2020 : “Probabilistic Load Forecasting Based on Optimal Quantile Regression Random Forest and Risk Severity Score”**

**by**

**Happy Aprillia, Hong-Tzer Yang and** **Chao-Ming Huang**

The authors are very grateful to the referees for their valuable comments upon this paper. In response to the concerns of the referees, we have made the following changes on the revised paper.

**Reviewer 1's comments:**

The paper "Probabilistic Load Forecasting Based on Optimal Quantile Regression Random Forest and Risk Severity Score" deals with a topic that is of interest for the TSG journal. It is overall well written and it brings sufficient novel contributions. There are some points that should however be clarified by the authors:

1) Feature mapping is an interesting contribution. The prior assumption on the normality of features should be discussed more in details, particularly considering that past load and weather features usually follow differentiated regimes. It appears that the authors considered this aspect by mapping features by the hour of the day, which is useful for some daily periodical loads but could be insufficient for different types of loads (e.g., industrial). This should be clearly stated in the body of the manuscript, with some comments on the weather regimes too.

2) The DWT decomposition requires the selection of the number of level of decomposition, and this is a hyper parameter of the model. How this should be selected? Also, why the WOA is used to optimally combine the individual decomposed signals if the simple sum of inverse DWT (almost) precisely reconstruct the original load? It appears that only the last detailed waveforms are removed in this way. Is the LSE calculated from the actual input load and the reconstructed load?

Please discuss this aspect, considering also the outcomes of [1] that presented a similar approach for the usage of wavelets in PLF.

3) It is mandatory to provide a theoretical justification to the usage of the RSS as a score to determine the skill of probabilistic forecasts. RSS is quite similar to the CWC which was discussed in [2] since it was not a strictly proper score [3]. It appears to me that RSS is not a strictly proper score, too, and therefore its usage for PLF evaluation is questioned. If the authors want to keep the RSS as a score to evaluate probabilistic forecasts, they must first provide a demonstration of the fact that RSS is strictly proper. Otherwise, RSS should be removed from the paper.

4) It is necessary to discuss the hyper parameter optimization of QRRFs. For example, how the number of trees (200) was selected?

5) Please provide more details regarding the weather predictions used ad inputs. Particularly, which is the forecast horizon of these NWPs? Do the paper targets day-ahead load forecasting? Please, discuss

[1] https://doi.org/10.1109/TSG.2019.2937072

[2] https://doi.org/10.1109/TSTE.2014.2323851

[3] https://doi.org/10.1198/016214506000001437

R1-1. Feature mapping is an interesting contribution. The prior assumption on the normality of features should be discussed more in details, particularly considering that past load and weather features usually follow differentiated regimes. It appears that the authors considered this aspect by mapping features by the hour of the day, which is useful for some daily periodical loads but could be insufficient for different types of loads (e.g., industrial). This should be clearly stated in the body of the manuscript, with some comments on the weather regimes too.

R1-2. The DWT decomposition requires the selection of the number of level of decomposition, and this is a hyper parameter of the model. How this should be selected? Also, why the WOA is used to optimally combine the individual decomposed signals if the simple sum of inverse DWT (almost) precisely reconstruct the original load? It appears that only the last detailed waveforms are removed in this way. Is the LSE calculated from the actual input load and the reconstructed load?

Please discuss this aspect, considering also the outcomes of [1] that presented a similar approach for the usage of wavelets in PLF.

R1-3. It is mandatory to provide a theoretical justification to the usage of the RSS as a score to determine the skill of probabilistic forecasts. RSS is quite similar to the CWC which was discussed in [2] since it was not a strictly proper score [3]. It appears to me that RSS is not a strictly proper score, too, and therefore its usage for PLF evaluation is questioned. If the authors want to keep the RSS as a score to evaluate probabilistic forecasts, they must first provide a demonstration of the fact that RSS is strictly proper. Otherwise, RSS should be removed from the paper.

R1-4. It is necessary to discuss the hyper parameter optimization of QRRFs. For example, how the number of trees (200) was selected?

R1-5. Please provide more details regarding the weather predictions used ad inputs. Particularly, which is the forecast horizon of these NWPs? Do the paper targets day-ahead load forecasting? Please, discuss

[1] <https://doi.org/10.1109/TSG.2019.2937072>

[2] <https://doi.org/10.1109/TSTE.2014.2323851>

[3] <https://doi.org/10.1198/016214506000001437>

**In response to Reviewer 1:**

R1-1. Feature mapping is an interesting contribution. The prior assumption on the normality of features should be discussed more in details, particularly considering that past load and weather features usually follow differentiated regimes. It appears that the authors considered this aspect by mapping features by the hour of the day, which is useful for some daily periodical loads but could be insufficient for different types of loads (e.g., industrial). This should be clearly stated in the body of the manuscript, with some comments on the weather regimes too.

**For R1-1:**  As the reviewer suggestion,

|  |
| --- |
| S |

R1-2. The DWT decomposition requires the selection of the number of level of decomposition, and this is a hyper parameter of the model. How this should be selected? Also, why the WOA is used to optimally combine the individual decomposed signals if the simple sum of inverse DWT (almost) precisely reconstruct the original load? It appears that only the last detailed waveforms are removed in this way. Is the LSE calculated from the actual input load and the reconstructed load?

Please discuss this aspect, considering also the outcomes of [1] that presented a similar approach for the usage of wavelets in PLF.

**For R1-2**  As the reviewer suggestion,

|  |
| --- |
| S |

R1-3. It is mandatory to provide a theoretical justification to the usage of the RSS as a score to determine the skill of probabilistic forecasts. RSS is quite similar to the CWC which was discussed in [2] since it was not a strictly proper score [3]. It appears to me that RSS is not a strictly proper score, too, and therefore its usage for PLF evaluation is questioned. If the authors want to keep the RSS as a score to evaluate probabilistic forecasts, they must first provide a demonstration of the fact that RSS is strictly proper. Otherwise, RSS should be removed from the paper.

**For R1-3:**  As the reviewer suggestion,

|  |
| --- |
| S |

R1-5. It is necessary to discuss the hyperparameter optimization of QRRFs. For example, how the number of trees (200) was selected?

**For R1-4:**  As the reviewer suggestion,

|  |
| --- |
| S |

R1-5. Please provide more details regarding the weather predictions used ad inputs. Particularly, which is the forecast horizon of these NWPs? Do the paper targets day-ahead load forecasting? Please, discuss

[1] <https://doi.org/10.1109/TSG.2019.2937072>

[2] <https://doi.org/10.1109/TSTE.2014.2323851>

[3] <https://doi.org/10.1198/016214506000001437>

**For R1-5:**  As the reviewer suggestion,

|  |
| --- |
| S |

**---------------------------------------------------------------------------------------------------------------------------------**

**Reviewer 2's comments:**

R2-1. The authors propose a model probability load forecasting that uses the quantile regression random forest technique, probability map, and risk severity score to obtain the actual pictorial of the outcome risk of the load demand profile.

The work is technically well written, but I still have some questions:

1) The most relevant input is obtained using correlation analysis, what is the initial input? (preconditioned historical load data and weather information). And what is the "most relevant" input selected?

2) The intervals prediction is produced with an accuracy profile varying by variances of 3%, 5%, and 10%. Why do the authors comment only on the results obtained for 3% accuracy?

3) The Table 2 shows the results of the proposed model and benchmarks, and are used as "CP", "NEW", "RSS" and "WS" scores. Why do the authors only comment on the "RSS" and "WS" scores?

4) The results obtained should be better explained. What would be the best results? Based on what? Why are the results obtained by the authors so different from the benchmark results?

5) It would be interesting if the authors improved the conclusion. Make it clear why the proposed model is reliable and applicable in the real system.

R2-1. The most relevant input is obtained using correlation analysis, what is the initial input? (preconditioned historical load data and weather information). And what is the "most relevant" input selected?

R2-2. The intervals prediction is produced with an accuracy profile varying by variances of 3%, 5%, and 10%. Why do the authors comment only on the results obtained for 3% accuracy?

R2-3. The Table 2 shows the results of the proposed model and benchmarks, and are used as "CP", "NEW", "RSS" and "WS" scores. Why do the authors only comment on the "RSS" and "WS" scores?

R2-4. The results obtained should be better explained. What would be the best results? Based on what? Why are the results obtained by the authors so different from the benchmark results?

R2-5. It would be interesting if the authors improved the conclusion. Make it clear why the proposed model is reliable and applicable in the real system.

**In response to Reviewer 2:**

**R2-1** The most relevant input is obtained using correlation analysis, what is the initial input? (preconditioned historical load data and weather information). And what is the "most relevant" input selected?

**For R2-1:** answer

|  |
| --- |
| answer |

**R2-2** The intervals prediction is produced with an accuracy profile varying by variances of 3%, 5%, and 10%. Why do the authors comment only on the results obtained for 3% accuracy?

**For R2-2:** answer

|  |
| --- |
| answer |

**R2-3** The intervals prediction is produced with an accuracy profile varying by variances of 3%, 5%, and 10%. Why do the authors comment only on the results obtained for 3% accuracy?

**For R2-3:** answer

|  |
| --- |
| answer |

**R2-4** The results obtained should be better explained. What would be the best results? Based on what? Why are the results obtained by the authors so different from the benchmark results?

**For R2-4:** answer

|  |
| --- |
| answer |

**R2-5** It would be interesting if the authors improved the conclusion. Make it clear why the proposed model is reliable and applicable in the real system.

**For R2-5:** answer

|  |
| --- |
| answer |

---------------------------------------------------------------------------------------------------------------------------------

**Reviewer 3's comments:**

I read this paper with a lot of interest. It presents a probabilistic method for short term load forecasting by using intervals. The method is not new, but it has been used in several applications; the advantage of the authors is that they recognize that and they provide all the relevant references. However, the application to short load forecasting is the novelty of the paper. Thus, I find that the paper has merit but there are some points that need to be addressed prior to being published. In particular, I have the following comments that need to be addressed (or responded):

- Why is the Winkler score used since there the RSS? What does the reader gain by the Winkler score?

- The sentence “The maximum and minimum uncertainty of the prediction can be observed, which facilitates decision making due to the clarity of analysis” needs to be supported by a reference.

- The paper fails to provide a review of learning probabilistic models in short term load forecasting. For instance the following papers should have been included:

o Alamaniotis, M., & Tsoukalas, L. H. (2017, September). Multi-kernel assimilation for prediction intervals in nodal short term load forecasting. In 2017 19th International Conference on Intelligent System Application to Power Systems (ISAP) (pp. 1-6). IEEE.

o Mori, H., & Ohmi, M. (2005, November). Probabilistic short-term load forecasting with Gaussian processes. In Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems (pp. 6-pp). IEEE.

- This sentence “Unlike the deterministic method, the probabilistic load forecasting does not rely on a specific prediction point and instead relies on a pair of intervals between the actual load 𝑌” is not clear. What are the pair of intervals?

- What is Mt and Np in equation (3)?

- Why are the weights of the NAW and CP are set as 0.5 and 0.5? is there a logic behind it?

- The proposed method is applied on four weeks taken from the four different seasons. That is the right approach and indeed it shows how the method performs on load dynamics driven by season. However, the validation and testing will be completely convincing to the reader only if the method is applied on special days. Since the authors used the data from USA (NE ISO), then I would like to see the performance on load prediction for days like thanksgiving, black Friday and new year and Christmas day. The load dynamics are different for those days and therefore I would love to see how the method performs for these days.

- The authors do not state how are their training dataset created. They have data only from two years 2016-2017 and therefore, I would like to see how are their training and validation data created? How many points are contained in the training data?

- Figure 8 is difficult to read. It appears to be blurred. It needs to be made with higher resolution or make it somehow clearer (perhaps save it bmp file and insert it in the text?).

- Table I is extended beyond the column boundaries.

- In figure 7 the 3% variance is very difficult to see. I would suggest authors that instead of colors use different type of curves. The latter is useful to readers that print out the paper in black and white.

- Table II is difficult to follow. I suggest that authors use bar-graphs where the reader is easier to read and understand. With bar graphs is also easier to understand the comparison among algorithms.

- The first column in Table I is entitled “method” but its contents refer to seasons.

- In section II, it is not very clear what do the authors mean by stating “multiple scenarios”?

- Can the authors provide the average variation per season? This metric would allow us to understand how volatile are the datasets and therefore how difficult is the validation and testing.

Overall, the paper has merit but also has several flaws. There is need to expand the testing datasets to make the paper convincing over the validity of the method.

R3-1. Why is the Winkler score used since there the RSS? What does the reader gain by the Winkler score?

R3-2. The sentence “The maximum and minimum uncertainty of the prediction can be observed, which facilitates decision making due to the clarity of analysis” needs to be supported by a reference.

R3-3. The paper fails to provide a review of learning probabilistic models in short term load forecasting. For instance, the following papers should have been included:

* Alamaniotis, M., & Tsoukalas, L. H. (2017, September). Multi-kernel assimilation for prediction intervals in nodal short term load forecasting. In 2017 19th International Conference on Intelligent System Application to Power Systems (ISAP) (pp. 1-6). IEEE.
* Mori, H., & Ohmi, M. (2005, November). Probabilistic short-term load forecasting with Gaussian processes. In Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems (pp. 6-pp). IEEE.

R3-4. This sentence “Unlike the deterministic method, the probabilistic load forecasting does not rely on a specific prediction point and instead relies on a pair of intervals between the actual load 𝑌” is not clear. What are the pair of intervals?

R3-5. What is Mt and Np in equation (3)?

R3-6. Why are the weights of the NAW and CP are set as 0.5 and 0.5? is there a logic behind it?

R3-7. The proposed method is applied on four weeks taken from the four different seasons. That is the right approach and indeed it shows how the method performs on load dynamics driven by season. However, the validation and testing will be completely convincing to the reader only if the method is applied on special days. Since the authors used the data from USA (NE ISO), then I would like to see the performance on load prediction for days like thanksgiving, black Friday and new year and Christmas day. The load dynamics are different for those days and therefore I would love to see how the method performs for these days.

R3-8. The authors do not state how are their training dataset created. They have data only from two years 2016-2017 and therefore, I would like to see how are their training and validation data created? How many points are contained in the training data?

R3-9. Figure 8 is difficult to read. It appears to be blurred. It needs to be made with higher resolution or make it somehow clearer (perhaps save it bmp file and insert it in the text?)

R3-10. Table I is extended beyond the column boundaries.

R3-11. In figure 7 the 3% variance is very difficult to see. I would suggest authors that instead of colors use different type of curves. The latter is useful to readers that print out the paper in black and white.

R3-12. Table II is difficult to follow. I suggest that authors use bar-graphs where the reader is easier to read and understand. With bar graphs is also easier to understand the comparison among algorithms.

R3-13. The first column in Table I is entitled “method” but its contents refer to seasons.

R3-14. In section II, it is not very clear what do the authors mean by stating “multiple scenarios”?

R3-15. Can the authors provide the average variation per season? This metric would allow us to understand how volatile are the datasets and therefore how difficult is the validation and testing.

R3-16. Overall, the paper has merit but also has several flaws. There is need to expand the testing datasets to make the paper convincing over the validity of the method.

**In response to Reviewer 3:**

**R3-1** Why is the Winkler score used since there the RSS? What does the reader gain by the Winkler score?

**For R3-1:** answer

|  |
| --- |
| answer |

**R3-2** The sentence “The maximum and minimum uncertainty of the prediction can be observed, which facilitates decision making due to the clarity of analysis” needs to be supported by a reference.

**For R3-2:** answer

|  |
| --- |
| answer |

**R3-3** The paper fails to provide a review of learning probabilistic models in short term load forecasting. For instance, the following papers should have been included:

• Alamaniotis, M., & Tsoukalas, L. H. (2017, September). Multi-kernel assimilation for prediction intervals in nodal short term load forecasting. In 2017 19th International Conference on Intelligent System Application to Power Systems (ISAP) (pp. 1-6). IEEE.

• Mori, H., & Ohmi, M. (2005, November). Probabilistic short-term load forecasting with Gaussian processes. In Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems (pp. 6-pp). IEEE.

**For R3-3:** answer

|  |
| --- |
| answer |

**R3-4** This sentence “Unlike the deterministic method, the probabilistic load forecasting does not rely on a specific prediction point and instead relies on a pair of intervals between the actual load 𝑌” is not clear. What are the pair of intervals?

**For R3-4:** answer

|  |
| --- |
| answer |

**R3-5** What is Mt and Np in equation (3)?

**For R3-5:** answer

|  |
| --- |
| answer |

**R3-6** Why are the weights of the NAW and CP are set as 0.5 and 0.5? is there a logic behind it?

**For R3-6:** answer

|  |
| --- |
| answer |

**R3-7** The proposed method is applied on four weeks taken from the four different seasons. That is the right approach and indeed it shows how the method performs on load dynamics driven by season. However, the validation and testing will be completely convincing to the reader only if the method is applied on special days. Since the authors used the data from USA (NE ISO), then I would like to see the performance on load prediction for days like thanksgiving, black Friday and new year and Christmas day. The load dynamics are different for those days and therefore I would love to see how the method performs for these days.

**For R3-7:** answer

|  |
| --- |
| answer |

**R3-8** The authors do not state how are their training dataset created. They have data only from two years 2016-2017 and therefore, I would like to see how are their training and validation data created? How many points are contained in the training data?

**For R3-8:** answer

|  |
| --- |
| answer |

**R3-9** Figure 8 is difficult to read. It appears to be blurred. It needs to be made with higher resolution or make it somehow clearer (perhaps save it bmp file and insert it in the text?)

**For R3-9:** answer

|  |
| --- |
| answer |

**R3-10** Table I is extended beyond the column boundaries.

**For R3-10:** answer

|  |
| --- |
| answer |

**R3-11** In figure 7 the 3% variance is very difficult to see. I would suggest authors that instead of colors use different type of curves. The latter is useful to readers that print out the paper in black and white.

**For R3-11:** answer

|  |
| --- |
| answer |

**R3-12** Table II is difficult to follow. I suggest that authors use bar-graphs where the reader is easier to read and understand. With bar graphs is also easier to understand the comparison among algorithms.

**For R3-12:** answer

|  |
| --- |
| answer |

**R3-13** The first column in Table I is entitled “method” but its contents refer to seasons.

**For R3-13:** answer

|  |
| --- |
| answer |

**R3-14** In section II, it is not very clear what do the authors mean by stating “multiple scenarios”?

**For R3-14:** answer

|  |
| --- |
| answer |

**R3-15** Can the authors provide the average variation per season? This metric would allow us to understand how volatile are the datasets and therefore how difficult is the validation and testing.

**For R3-15:** answer

|  |
| --- |
| answer |

**R3-16** Overall, the paper has merit but also has several flaws. There is need to expand the testing datasets to make the paper convincing over the validity of the method.

**For R3-16:** answer

|  |
| --- |
| answer |

---------------------------------------------------------------------------------------------------------------------------------

**Reviewer 4's comments:**

R4-1. The manuscript proposes a probabilistic load forecasting method based on optimal quantile regression and compares it with other techniques available in literature. In general terms, the article follows a classic structure, where the proposed method is compared against a number of alternatives, it is clear and well-written.

However, in the opinion of this reviewer, the proposed method needs still need to compared with other methods which are part of the state of the art, in terms of time series forecasting. Particularly, authors need to compare the proposed approach with the results that could be obtained using time-varying Markov Chains, which are particularly suitable for multi-modal probability distributions. The main reason behind this request is that the concept of quantile is suitable for uni-modal distributions, but fails to characterize the complexity of the process when there are two or more likely scenarios that need to be considered in the forecast. Also, performance metrics require to include those generally used in the domain of Information and Bayesian theory, where the resulting prior probability distribution is characterized in terms of the quality of the information that is included during the training procedure. For example, how does the quality of the uncertainty representation degrade as time evolves? How well is represented the fact that uncertainty grows as you move forward in the future? Are all uncertainty sources well represented and included in the analysis?

The paper presents a method and analyzes its performance based on a simple numeric evaluation of results. This analysis requires an in-depth theoretical study, where it is clearly explained how the method is able to characterize uncertainty sources adequately. That makes the difference between a probability-based approach and a statistical method. In my humble opinion, this method can hardly be categorized as "probabilistic".

In this regard, this paper cannot be published before that analysis is conducted.

R4-1. However, in the opinion of this reviewer, the proposed method needs still need to compared with other methods which are part of the state of the art, in terms of time series forecasting. Particularly, authors need to compare the proposed approach with the results that could be obtained using time-varying Markov Chains, which are particularly suitable for multi-modal probability distributions. The main reason behind this request is that the concept of quantile is suitable for uni-modal distributions, but fails to characterize the complexity of the process when there are two or more likely scenarios that need to be considered in the forecast.

R4-2. Also, performance metrics require to include those generally used in the domain of Information and Bayesian theory, where the resulting prior probability distribution is characterized in terms of the quality of the information that is included during the training procedure. For example, how does the quality of the uncertainty representation degrade as time evolves? How well is represented the fact that uncertainty grows as you move forward in the future? Are all uncertainty sources well represente and included in the analysis?

R4-3. The paper presents a method and analyzes its performance based on a simple numeric evaluation of results. This analysis requires an in-depth theoretical study, where it is clearly explained how the method is able to characterize uncertainty sources adequately. That makes the difference between a probability-based approach and a statistical method. In my humble opinion, this method can hardly be categorized as "probabilistic".

**In response to Reviewer 4:**

**R4-1** However, in the opinion of this reviewer, the proposed method needs still need to compared with other methods which are part of the state of the art, in terms of time series forecasting. Particularly, authors need to compare the proposed approach with the results that could be obtained using time-varying Markov Chains, which are particularly suitable for multi-modal probability distributions. The main reason behind this request is that the concept of quantile is suitable for uni-modal distributions, but fails to characterize the complexity of the process when there are two or more likely scenarios that need to be considered in the forecast.

**For R4-1:** answer

|  |
| --- |
| answer |

**R4-2** Also, performance metrics require to include those generally used in the domain of Information and Bayesian theory, where the resulting prior probability distribution is characterized in terms of the quality of the information that is included during the training procedure. For example, how does the quality of the uncertainty representation degrade as time evolves? How well is represented the fact that uncertainty grows as you move forward in the future? Are all uncertainty sources well represented and included in the analysis?

**For R4-2:**  answer

**For R4-2:** answer

|  |
| --- |
| answer |

**R4-3** The paper presents a method and analyzes its performance based on a simple numeric evaluation of results. This analysis requires an in-depth theoretical study, where it is clearly explained how the method is able to characterize uncertainty sources adequately. That makes the difference between a probability-based approach and a statistical method. In my humble opinion, this method can hardly be categorized as "probabilistic".

**For R4-3:** answer

|  |
| --- |
| answer |

---------------------------------------------------------------------------------------------------------------------------------

**Reviewer 5's comments:**

R5-1. 1、 The notation is not clear enough, which should be further enhanced and carefully reviewed.

2、 For readers to quickly catch your contribution, it would be better to highlight major difficulties and challenges, and your original achievements to overcome them, in a clearer way in abstract and introduction.

3、 In training stage, how to modeled as the input variables and actual load a normal distribution? How to achieve the conditional probability? The process should be detailed description.

4、 In testing stage, actual Load are used to build probability map of input variables. Whether these actual Load are forecasting targets?

5、 In training stage, how to build the probability map? According to the formula (12), the probability map is a kind of discrete mapping. However, the input variables and actual load are continuous variable. This conflict is how to resolved.

6、 It is mentioned in p.3 that the WOA and DWT method is adopted before feeding the input into the QRRF. What are other feasible alternatives? What are the advantages of adopting this particular method over others in this case? How will this affect the results? As the second claimed contribution, what is the foothold of its innovation? The authors should provide more details on this.

7、 In Section III, the point forecasting model is not presented. In part B, WOA-DWT and QRRF are used to produce point forecasting? How to produce probability forecasting is not mentioned. The authors should provide more details on this.

8、 In Section III.C, RSS is proposed to evaluate prediction interval. What is the necessity of putting forward this index?

9、 It is mentioned in p.5 that Winkler score is adopted to evaluate the prediction interval resulting from the proposed probabilistic load forecasting and to demonstrate the effectiveness of the RSS. What are other feasible alternatives? What are the advantages of adopting this particular evaluation index over others in this case? How will this affect the results? The authors should provide more details on this.

10、What are the input variables for the ISO-NE dataset?

11、 A lower RSS and WS can be assigned as the better prediction interval. However, in case study, for the 3%, 5%, and 10% variances, these two index present opposite result. The authors should provide more details on this.

12、It is mentioned in p.6 that three kinds of benchmark methods are adopted as benchmark for comparison. What are the other feasible alternatives? What are the advantages of adopting these particular methods over others in this case? How will this affect the results? More details should be furnished.

13、Some key parameters are not mentioned. The rationale on the choice of the particular set of parameters should be explained with more details. Have the authors experimented with other sets of values? What are the sensitivities of these parameters on the results?

R5-1. The notation is not clear enough, which should be further enhanced and carefully reviewed.

R5-2. For readers to quickly catch your contribution, it would be better to highlight major difficulties and challenges, and your original achievements to overcome them, in a clearer way in abstract and introduction.

R5-3. In training stage, how to modeled as the input variables and actual load a normal distribution? How to achieve the conditional probability? The process should be detailed description.

R5-4. In testing stage, actual Load are used to build probability map of input variables. Whether these actual Load are forecasting targets?

R5-5. In training stage, how to build the probability map? According to the formula (12), the probability map is a kind of discrete mapping. However, the input variables and actual load are continuous variable. This conflict is how to resolved.

R5-6. It is mentioned in p.3 that the WOA and DWT method is adopted before feeding the input into the QRRF. What are other feasible alternatives? What are the advantages of adopting this particular method over others in this case? How will this affect the results? As the second claimed contribution, what is the foothold of its innovation? The authors should provide more details on this.

R5-7. In Section III, the point forecasting model is not presented. In part B, WOA-DWT and QRRF are used to produce point forecasting? How to produce probability forecasting is not mentioned. The authors should provide more details on this.

R5-8. In Section III.C, RSS is proposed to evaluate prediction interval. What is the necessity of putting forward this index?

R5-9. It is mentioned in p.5 that Winkler score is adopted to evaluate the prediction interval resulting from the proposed probabilistic load forecasting and to demonstrate the effectiveness of the RSS. What are other feasible alternatives? What are the advantages of adopting this particular evaluation index over others in this case? How will this affect the results? The authors should provide more details on this.

R5-10. What are the input variables for the ISO-NE dataset?

R5-11. A lower RSS and WS can be assigned as the better prediction interval. However, in case study, for the 3%, 5%, and 10% variances, these two index present opposite result. The authors should provide more details on this.

R5-12. It is mentioned in p.6 that three kinds of benchmark methods are adopted as benchmark for comparison. What are the other feasible alternatives? What are the advantages of adopting these particular methods over others in this case? How will this affect the results? More details should be furnished.

R5-13. Some key parameters are not mentioned. The rationale on the choice of the particular set of parameters should be explained with more details. Have the authors experimented with other sets of values? What are the sensitivities of these parameters on the results?

**In response to Reviewer 5:**

**R5-1** The notation is not clear enough, which should be further enhanced and carefully reviewed.

**For R5-1:**  answer

**For R5-1:** answer

|  |
| --- |
| answer |

**R5-2** For readers to quickly catch your contribution, it would be better to highlight major difficulties and challenges, and your original achievements to overcome them, in a clearer way in abstract and introduction.

**For R5-2:**  answer

**For R5-2:** answer

|  |
| --- |
| answer |

**R5-3** In training stage, how to modeled as the input variables and actual load a normal distribution? How to achieve the conditional probability? The process should be detailed description.

**For R5-3:**  answer

**For R5-3:** answer

|  |
| --- |
| answer |

**R5-4** In testing stage, actual Load are used to build probability map of input variables. Whether these actual Load are forecasting targets?

**For R5-4:**  answer

**For R5-4:** answer

|  |
| --- |
| answer |

**R5-5** In training stage, how to build the probability map? According to the formula (12), the probability map is a kind of discrete mapping. However, the input variables and actual load are continuous variable. This conflict is how to resolved.

**For R5-5:**  answer

**For R5-5:** answer

|  |
| --- |
| answer |

**R5-6** It is mentioned in p.3 that the WOA and DWT method is adopted before feeding the input into the QRRF. What are other feasible alternatives? What are the advantages of adopting this particular method over others in this case? How will this affect the results? As the second claimed contribution, what is the foothold of its innovation? The authors should provide more details on this.

**For R5-6:** answer

|  |
| --- |
| answer |

**R5-7** In Section III, the point forecasting model is not presented. In part B, WOA-DWT and QRRF are used to produce point forecasting? How to produce probability forecasting is not mentioned. The authors should provide more details on this.

**For R5-7:** answer

|  |
| --- |
| answer |

**R5-8** In Section III.C, RSS is proposed to evaluate prediction interval. What is the necessity of putting forward this index?

**For R5-8:** answer

|  |
| --- |
| answer |

**R5-9** It is mentioned in p.5 that Winkler score is adopted to evaluate the prediction interval resulting from the proposed probabilistic load forecasting and to demonstrate the effectiveness of the RSS. What are other feasible alternatives? What are the advantages of adopting this particular evaluation index over others in this case? How will this affect the results? The authors should provide more details on this.

**For R5-9:** answer

|  |
| --- |
| answer |

**R5-10** What are the input variables for the ISO-NE dataset?

**For R5-10:** answer

|  |
| --- |
| answer |

**R5-11** A lower RSS and WS can be assigned as the better prediction interval. However, in case study, for the 3%, 5%, and 10% variances, these two index present opposite result. The authors should provide more details on this.

**For R5-11:** answer

|  |
| --- |
| answer |

**R5-12** It is mentioned in p.6 that three kinds of benchmark methods are adopted as benchmark for comparison. What are the other feasible alternatives? What are the advantages of adopting these particular methods over others in this case? How will this affect the results? More details should be furnished.

**For R5-12:** answer

|  |
| --- |
| answer |

**R5-13** Some key parameters are not mentioned. The rationale on the choice of the particular set of parameters should be explained with more details. Have the authors experimented with other sets of values? What are the sensitivities of these parameters on the results?

**For R5-13:** answer

|  |
| --- |
| answer |