Review Simulating run-to-failure SCADA time series to enhance wind turbine fault detection and prognosis.

The paper concerns the detection/prediction of faults and remaining useful life (RUL) estimation for wind turbines. It presents a methodology where synthetic data, produced by a cGAN algorithm, is used to improve the accuracy of failure detection and RUL algorithms. The methodology is validated on data from a wind farm with a single failure and one near failure (maintenance).

The problem statement is very relevant. Failure prediction and RUL on wind turbines is an important research topic. This paper can be an interesting and valuable contribution. This means that in my opinion the topic is in scope of the WES journal.

The paper is well written, reads fluent, and is to the point. The 18 pages of text (excluding the references) describes well the problem statement. The methodology is relatively well discussed, however, certain parts are unclear or missing (see remarks part). The results are well presented. The conclusion is concise and to the point but in my opinion somewhat too strong in its assertions (see remarks part). The abstract is concise and to the point and describes the content of the paper well. The figures that are added to the paper are useful to understand the text. The figures are also clear.

The paper shows that when synthetic data is used using a cGAN the number of misclassified labels (unhealthy, healthy) decreases by 84% on average. The paper indicates how the synthetic data methodology improves the performance in this case.

I do have some remarks/questions that in my opinion need clarification in the paper:

On p.4 it is said “the weights of the f_R and f_Std are experimentally set to 3 to balance the four terms”. How was this done? Which data was used for the experiment?

On p.6-8 the methodology is discussed for generating the synthetic SCADA signals using cGAN. Can you state more clearly what the hyperparameter values were for this algorithm? For example the number of layers in the networks, the number of neurons, ...
On p.12 it is mentioned that the learning rate of the Adam optimizer is 0.0005. How were the hyperparameter values selected? Was hyperparameter tuning used? If so, how was this done? Which data was used for this? If no hyperparameter tuning was done, then where do the hyperparameter values come from? From literature? If so, is it not surprising that those parameters work well on the case discussed in the paper? Can you please discuss this?

On p.7 it is said that the window length w is experimentally set to 10. How was this experimentally done? Which data was used for the experiment? Which metric was used to decide which window length is optimal?

On p.9 it is said “Each turbine has a diameter of 100 m …”. Although it is clear from the context that “rotor diameter” is meant, it might be useful to explicitly specify it is the rotor diameter.

On p.9 it is said that WT9 is selected as validation dataset. Why was this wind turbine selected? How was it selected? This dataset is used to set the detection threshold. It is known that the component temperatures between different wind turbines can be structurally different even if they are healthy. It is surprising that this did not have an impact on the number of false positives in the test dataset. Did you notice differences between the different turbines? Please discuss this.

On p.10 you say the following: “… the minority (faulty) class is oversampled using the SMOTEN method …”. Should it not be SMOTE? How was SMOTE applied during the training of the model? On which data? Please discuss this.

As a baseline a classification model is trained on data from a wind turbine that experienced a gearbox failure (p.10). No more information on the failure type is given. 12 months of data preceding the failure are used for training the model. Data less than 1 month before the failure is labeled as unhealthy, the remaining 11 months are considered healthy. It is not explained in the paper how this decision was taken. Why 1 month and not more or less data? It might be good to discuss this in the paper.

It might also be good to add a bit more information on how this classification model was trained. Was a train-validation split used? If so, how much data was assigned to training and how much to validation? On p.10 the architecture is described. It is stated that this architecture is a good trade-off between performance and computational burden. How was this decided? Was hyperparameter tuning performed? If so, how was it done, on which data? If no hyperparameter tuning was done, how were the values for the hyperparameters found? Please discuss this.

Furthermore, by using only one failure for training, does this not risk overly specializing the model to the degradation pattern of this single failure? My experience is that failures, even if they are of the same type, show themselves often quite differently in the data. There tends to be large variance in the degradation patterns. How does this method handle this? How do you guarantee that the model does not just memorized this degradation pattern (or certain properties of it)? Have you tested it on other failure types? The fact that there are no false positives for the “healthy” turbines is surprising. Gearbox temperatures are most likely influenced by many environmental conditions, which are not always easy to measure and use in a model. So I would expect there to be more noise on the results. Please discuss this.

Figure 7 and the explanation on p.12 indicates that an anomaly zone was identified for WT6, and that this was most likely associated with a maintenance. An analysis of the pattern shows that there was an initial jump of the fault index, then it decreased sharply

and then it stayed at a lower but elevated level. If it identified damage to the gearbox, why do we not see an increasing trend in the fault index over time? What causes the initial jump, and subsequent sharp decrease? It might be useful to add some discussion of this to the paper.

On p.14-15, it is explained how 4 trends are used for the synthetic HIs. These are all based on characteristics of the failure for WT8. Does this method not risk overly specializing the model to this single failure? Does this not mean that the good results achieved are limited to this single failure? What can be expected in conditions with multiple failures, how would the methodology be applied? What will be the impact of this? It might be useful to discuss this.

On p.16 the results of the RUL estimation are discussed. A second order polynomial is fit on the HI to predict the RUL up to the detection point. How was the decision taken to use a second order polynomial? What was the procedure? Was the decision taken by looking at the shape of the HIs? If this is the case don't you run the risk of overfitting this specific case? Is using a second order polynomial still valid when testing on other failures (degradation patterns)? Again the variance in the degradation patterns plays a role here. Please discuss this.

In my opinion the paper results can be seen as a proof of concept of a methodology. Quite some assumptions (a.o. how a degradation signal looks like, …) are made during the construction of the pipeline. To know how well it would perform in general, a larger analysis is required on more failures (of different types). It might be useful to add this point more clearly to the conclusion of the paper.

It might also be useful to add a schematic overview of which data was used by which part of the pipeline.

Conclusion: This paper is in scope of WES. It contains relevant research, and is in my opinion after addition of the discussions good for publication.