Challenges in Using Operational Data for Reliable Wind Turbine Condition Monitoring

Jannis Tautz-Weinert, Simon J Watson
Centre for Renewable Energy Systems Technology, Wolfson School, Loughborough University
Loughborough, Leicestershire, United Kingdom

ABSTRACT

Operational data of wind turbines recorded by the Supervisory Control And Data Acquisition (SCADA) system originally intended only for operation and performance monitoring show promise also for assessing the health of the turbines. Using these data for monitoring mechanical components, in particular the drivetrain subassembly with gearbox and bearings, has recently been investigated with multiple techniques. In this paper the advantages and drawbacks of suggested approaches as well as general challenges and limitations are discussed focusing on automated and farm-wide condition monitoring.

KEY WORDS: Wind Turbine; Condition Monitoring; SCADA; Drivetrain; Machine Learning.

INTRODUCTION

Optimisation of maintenance is essential to further reduce the costs of offshore wind energy, where accessibility is restricted by weather conditions and the availability of transport vessels. Advanced maintenance strategies involve condition based decision-making while trying to predict the future maintenance needs before critical failures with significant downtimes occur. Continuous and reliable information of the condition of the different subassemblies and parts of the wind turbine are needed for effective prognosis of ongoing degradation and estimation of remaining life of critical parts.

Supervisory Control And Data Acquisition (SCADA) data have gained more attention in the last five years as they are usually available without any additional expense in contrast to dedicated condition monitoring systems which can cost approx. £14,000 per turbine (Yang et al., 2014). The operational data recorded in a SCADA system vary with the turbine type, but usually include at least wind speed, wind direction, yaw angle, pitch angle, active power, reactive power, generator current, generator speed, gearbox temperature, generator winding temperature and ambient temperature. Comparing parameters over time and in relation to the operational level has helped to identify changes in the behaviour related to developing failure (Wiggelinkhuizen et al., 2008; Feng et al., 2013). Based on that, the main idea has been the modelling of signals, mainly temperatures, assuming normal conditions and revealing problems via comparing modelled and measured temperatures. The focus of research has been on data-driven training of algorithms and machine learning tools adapted from computer science have been proposed, e.g. artificial neural networks (Garcia et al., 2006; Zaher et al., 2009; Bangalore and Tjernberg, 2015; Sun et al., 2016), adaptive-neuro fuzzy inference systems (Schlechtingen et al., 2013), nonlinear state estimation techniques (Wang and Infield, 2012) or multivariate adaptive regression splines (Tan and Zhang, 2016). An overview of the progress in the area of condition monitoring with operational data can be found in a recent review of the authors (Tautz-Weinert and Watson, 2016a).

Most publications on condition monitoring with operational data consist of a proposal for a new technique and a demonstration using one case study, whereas difficulties and challenges are rarely discussed. Yang et al. (2014) highlighted in their review of the current challenges in wind turbine condition monitoring that the sampling resolution of SCADA data is too low to monitor all aspects of a wind turbine and doubted the usefulness of SCADA monitoring in terms of early detection. The authors suggested the integration of SCADA-based monitoring in condition monitoring systems, however. Dienst and Beseler (2016) shared their lessons learned from monitoring of an offshore wind farm with operational data indicating e.g. that finding training data without errors is difficult, 2% of sensors are malfunctioning at any given time, using multiple models to predict a signal are beneficial and anomalies in models can indicate a defect but also unrepresentative training.

This work addresses the challenges in using operational data for wind turbine monitoring with the approach of normal behaviour modelling of temperatures based on experiences with real data from four wind farms. Drawbacks of the individual techniques are discussed and general challenges highlighted regarding data quality, pre-processing, input selection and alarm generation.

In the next section, the basic idea of normal behaviour modelling is introduced. The subsequent and third section summarises the properties of the data used. The fourth and main section addresses the challenges if failures are to be found retrospectively, whereas the fifth section gives a brief outlook if such an approach is used on-line. In the last section this work is concluded by summarising the key problems to be solved.

MONITORING BY NORMAL BEHAVIOUR MODELLING

Normal behaviour modelling is a way of building a virtual clone of a system which always represents the healthy state. The model generates a time series of the target signal, which can be compared with the measured signal to detect anomalies. Due to the complexity of wind turbine systems such models cannot be built analytically, but are data-driven. In a training period, where the turbine is assumed to operate normally, the relationship between input signals and the target signal is learnt by the algorithms.

Adequate target signals and corresponding inputs have to be selected to achieve a model which is useful in failure detection. Simple models predicting a signal with a sensor signal of the same type at the same location might help for monitoring the sensor itself. More advanced models can be used to monitor mechanical parts which are affected by wear: drivetrain bearings and gears. Wear will change the efficiency of a part and result in increased thermal losses which should become visible in the form of changed thermal behaviour (Feng et al., 2013). To monitor wear-related parts, temperature signals are commonly modelled with other temperatures of surrounding parts, signals describing the turbine's load level such as power output, electrical currents, rotational or wind speeds and/or signals representing the environmental background such as the nacelle or ambient temperature.

CASE STUDY

Records from four wind farms are used to highlight the challenges in condition monitoring with operational data. SCADA data are retrospectively analysed aiming to detect failures in advance. Although the turbines are from different manufacturers, all turbines are geared, variable speed and pitch controlled. The turbines cover the 1.5 MW and the 2-3 MW class. The investigated records range from only half a year to nearly five years and from 11 to 102 turbines in a farm. Reports of replacements are available for three of four farms. Although records from farm A are not supported by sufficient reports for failure detection analysis, normal behaviour modelling can be tested and compared based on the SCADA data. The key features of the data used are summarised in Table 1.

Table 1. Wind farm data used in this study

Farm	Location	Power (MW)	Number of turbines	Length of data (years)	Service report
A	USA	1.5	108	0.5	Stoppages only
B	UK	2-3	12	2.5	Stoppages and replace-ments
C	Europe	2-3	25	3.0	Replace-ments
D	Europe	2-3	11	4.7	Replace-ments

CHALLENGES IN RETROSPECTIVE ANALYSES

The challenges in using operational data for condition monitoring can be divided into: data quality, monitoring setup, proposed modelling techniques, comparing modelling techniques, modelling capabilities and alarm generation.

Data quality

Retrospective failure detection based on operational data is conducted with two main types of information: SCADA records available in a SQL database or spreadsheets and a service record in a spreadsheet.

SCADA data. Although signals in SCADA records are usually named, the labelling of the signals is not necessarily sufficient for clear identification of the sensor properties. As there is neither a common set of available signals nor a generally accepted taxonomy, different SCADA systems use different names and abbreviations. Although unambiguous signals like the power output, wind speed, blade pitch angle etc. are always easily identifiable, other signals require more details for complete identification. In particular, the location of temperature sensors is often insufficiently described. In the investigated data the labelling ranged from only numbering all temperature sensors (e.g. temperature 2, farm B), giving the name of the subassembly (e.g. gearbox temperature, farm A), specifying a part type in a subassembly (e.g. gearbox bearing temperature, farm C) to providing approx. location of the sensor at a part (e.g. gearbox bearing high speed shaft gearbox [vicinity / side], farm D). Even in the farm with the most detailed labelling, the locations are open to interpretation: e.g. there are two generator bearing sensors labelled 1 and 2 or oil temperatures are labelled basis, level 1 and 2. Detailed knowledge of the turbine configuration or a technical drawing including the sensor locations would certainly ease the analysis but has not been available for this work. Reasons can be found in insufficient documentation and confidentiality issues applicable to academic studies with commercial data.

Although missing, invalid and poorly processed data hinder the analysis, the most serious problems are caused by inconsistencies. Any change in the behaviour of a sensor might be interpreted as a change of the monitored part. In data from farm D, several changes of the maximum occurring values can been observed as shown for example in Fig. 1. Sensor specifications or detailed information about the operation are not available. It is assumed that this event could be caused by a sensor drift, unreported maintenance or a change in control and operation. An actual change of the performance of the monitored part without any interaction by the operator is unlikely due to the rapid change. Additionally, the temperatures in the illustrated example are lower after the step, i.e. the losses would be reduced which is contrary to the effects of wear. To allow analysis if inconsistencies occur, data should be split into windows without steps which are investigated separately. A systematic way of detecting steps is required for automated splitting and applying the training and testing procedure of normal behaviour modelling. Comparing monthly maximums and percentiles resulted in adequate detection of steps.

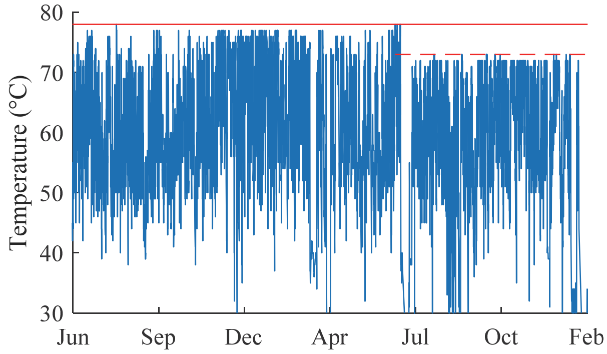

Fig. 1. Example of sensor inconsistency in a bearing temperature with a maximum temperature of 78° C before and 73° C after the step (farm D)

Service record. Insufficient documentation plays a major role if monitoring techniques are evaluated with real data. The service record consisted in the investigated case study of a list of stoppages in the best case (farm B). Comments were added only for major replacements or occasionally for other maintenance actions describing the reason for the stoppage time. Assumed reasons for replacements and interpretations of alarms, stoppages and inspections were generally missing. Accordingly, the list of replacements is not a list of failures. Replacements could have been done as preventative interventions or after a failure which had caused the turbine to stop. Additionally, the time of replacement is not necessarily the time of the failure or the detection of the failure. For the other investigated data, the failure record consisted only of a list of replacements (farm C and D) or was not available at all (farm A).

Although it can be assumed that the operator or service provider has always full access to all reports, the shortcoming of incomplete or incomprehensible service reports is widely acknowledged. Accordingly, service providers are currently focussing on the digitisation of reporting and implementation of procedures to improve the data quality e.g. by using mobile devices for documentation.

Monitoring techniques based on operational data have to be developed and tested with real data. It is very rare to get data of good quality and complete information in terms of turbine and sensor specifications or operation and maintenance reports. As this will similarly be true for industrial application, any modelling technique has to cope with incomplete information. However, the impact of data quality problems should be carefully considered when findings are generalised.

Monitoring setup

The detailed configuration of the monitoring setup consists of multiple choices in terms of the model architecture, input selection, pre-processing and training length.

Model architecture. A model using other signals to predict the target can be denoted as full signal reconstruction (FSRC) (Schlechtingen and Santos, 2010). Modelling could use the signal of the same time as the target or from previous time-steps to account for the inertia of the system. Using the latest history of the target itself could also be chosen to form an autoregressive model. If the history of the target is combined with other inputs, the model can be denoted as autoregressive with exogenous input (ARX). Although ARX models are more accurate in predicting the target, the prediction is likely to adapt to new behaviour, which might hinder failure detection.

Input selection. Selecting the inputs for modelling has commonly been done based on the physical understanding of the system also called domain knowledge (Schlechtingen and Santos, 2010; Wang and Infield, 2012; Bangalore and Tjernberg, 2015; Sun et al., 2016) or by correlation analyses between possible inputs and the target (Zaher et al., 2009; Tautz-Weinert and Watson, 2016b). Although most domain knowledge approaches have been based on the basic idea of the heat transfer in the drive train, the reasons for the manual choices of inputs have not been documented thoroughly. The limitation of possible inputs is additionally in contrast to the idea of using machine learning to find complex relationships. Using automated correlation analyses to build the model has the risk of selecting multiple similar inputs, e.g. generator currents 1-3.

The case studies show that e.g. for an ambient temperature a low correlation with a bearing temperature results in an exclusion as input, but less seasonal error is observed if the ambient temperature is selected as input. Selecting inputs only based on their correlation to the target is accordingly not necessarily the best option. Using all possible inputs and an algorithm to select inputs based on their relevance for accurate prediction has been proposed by Dienst and Beseler (2016) in applying the Least Absolute Shrinkage and Selection Operator (LASSO).

Pre-processing. Pre-processing of inputs should include a validity check to exclude data acquisition errors and time-steps with missing or erroneous data have to be removed completely. Additionally, scaling and lag removal might be necessary depending on the modelling technique and input selection (Schlechtingen and Santos, 2010). Focussing on data when the turbine is operating might ease the modelling and failure detection and can be implemented by filtering with a power threshold (Sun et al., 2016).

Training length. There is no consensus about the necessary training time under (assumed) healthy conditions. The proposed lengths range from 3 (Schlechtingen and Santos, 2010) to 14 months (Bach-Andersen et al., 2016). A length of one year is obviously beneficial to cover the full seasonal variation, but such a long training time is probably not always achievable. Other work has tried to concentrate data from a longer period in a shorter, representative training set (Wang and Infield, 2012; Bangalore and Tjernberg, 2014; Tan and Zhang, 2016), although there is not necessarily a benefit in terms of the modelling accuracy.

Tests in the case study reveal that the required training length depends on the turbine specific operation and behaviour. Even one month training results in acceptable accuracy for some turbines. Future work should address the sensitivity to the training length in more detail.

Proposed modelling techniques

Several models have been proposed for the required regression task of normal behaviour modelling.

Multi-linear regression (LIN). LIN trained by a least square algorithm is a simple way of modelling the system. Although this assumes linearity, the prediction can have a similar level of accuracy compared with more complex tools (Schlechtingen and Santos, 2010; Tautz-Weinert and Watson, 2016b). Slight improvements have been proposed such as allowing selected polynomial terms up to ninth order (Wilkinson et al., 2014), interactions, i.e. products of the inputs (Tautz-Weinert and Watson, 2016b) or added features such as squares, roots and logarithms (Dienst and Beseler, 2016).

Artificial Neural Networks (ANNs). ANNs have been widely applied to extend the modelling capabilities to non-linearity. A basic setup uses a feed-forward backpropagation network with one input, one hidden layer with a small number of neurons and one output layer with a single linear output. Past research has involved a range of configurations, though authors do not always describe in detail the set up used, Zaher et al. (2009) found that 3 neurons in one hidden layer provide the best results in an ARX approach to model the gearbox bearing. In contrast, Bangalore and Tjernberg (2014) use two hidden layers with 13 neurons in the first layer and one neuron in the second layer in a FSRC approach for gearbox monitoring. Sun et al. (2016) state that the number of neurons has to be selected for each turbine individually ranging from 2 to 10. L. Wang et al. (2016) claim that so called deep ANNs are better with three hidden layers of 100 neurons. Tan and Zhang (2016) highlight the difficulty in selecting a configuration when randomly varying the number of neurons and the type of transfer function and trying to select the best of 200 ANNs.

Tests in the case study show that using more neurons generally improves the accuracy, but there is no significant advantage of deep ANNs with three layers of 100 neurons. Varying the number of neurons and their transfer function randomly does not counterbalance the worse performance of some turbines if using a fixed configuration.

Adaptive Neuro-Fuzzy Inference Systems (ANFIS). ANFIS as a way of learning a fuzzy system with ANN approaches have been proposed by Schlechtingen, Santos and Achiche (2013). Two inputs were used per target with generalised normal distribution membership functions and hybrid gradient descent and least squares estimation learning. The main advantage over straight ANNs was given as the reduced training time.

Nonlinear State Estimation Technique (NSET). NSET has been proposed by Y. Wang and Infield (2012) as a way of modelling based on a state matrix and a weighting vector determined by a least square approach and a Euclidean distance operator. NSET includes the target signal in the state matrix and for determining the weighting vector. It is accordingly comparable to an ARX approach. To find a good compromise of better accuracy for more states and reasonable computational effort for fewer states, a data selection algorithm was proposed. The algorithm selects states, if they are less than the defined distance δ away from a regular grid of 100 sections of the normalised input. However, the algorithm allowed multiple states for the same grid point which resulted in a high numbers of states. Reproducing the approach in the case studies shows that it was impossible to get a similar number of states for different turbines with one selected δ. However, Guo, Infield and Yang (2012) defined the algorithm to select only one state per grid point. Testing this approach results in a dramatically lower, but more regular number of states for different turbines.

Multivariate Adaptive Regression Splines (MARS). MARS have been applied to wind turbine normal behaviour modelling by Tan and Zhang (2016) allowing a maximum of 21 basis functions. Each basis function can be a constant (for the intercept), a hinge function or a product of hinge functions.

Further well known techniques such as Gaussian Process and Support Vector Machine could also be used for the regression task. However, first case study results could not demonstrate any advantage in using these techniques (Tautz-Weinert and Watson, 2016b).

Comparing modelling techniques

The proposed modelling techniques and different input choices are compared with configurations as detailed in Table 2. A comparison of different modelling techniques should consider two main features: effort and accuracy.

The evaluation of the effort can be expressed in the simplicity of the model and the computational effort in training. The simplicity is here evaluated based on the subjective experience of implementing the technique in MATLAB 2015b. Computational effort is easily comparable in terms of the runtime. Example numbers are given using a common desktop PC (64-bit operating system with a four core CPU with 2.8 GHz clock rate and 32 GB memory).

Table 2. Modelling setups for comparison

Technique	Properties
General	- Data from farms A-D, pre-processed including non-operation filtering, turbines with known failures excluded. - Modelling target: Gearbox (bearing) temperature - Modelling inputs: a) 2 inputs, b) 3 inputs selected based on correlation c) power and rotational speed, d) power, rotational speed and ambient temperature - 3 months training, 3 months' blind testing
LIN	Linear terms and interactions.
ANN	FSRC feed-forward network with 20 neurons in hidden layer.
ANFIS	2 generalised normal distribution membership functions per input.
NSET	One state per grid point, $\delta = 0.001$.
MARS	Maximum of 21 basis functions.

There are internal MATLAB functions for all discussed methods, except MARS for which a toolbox is available online (Jekabsons, 2016) and NSET which has been implemented according to Wang and Infield (2012). LIN does not require any detailed configuration and is consequently the easiest method to implement. In contrast, settings have to be chosen for ANNs, ANFIS and MARS. The default settings and main approaches in the literature might be useable for configuring MARS and ANFIS, but in particular the choice of the architecture of ANNs appears to be surprisingly random.

The major advantage of linear models is the low computational effort required as shown in Table 3. Training of ANNs also requires relatively low computational effort with approx. 1-3 s per turbine, but training deep ANNs or repeating the training hundreds of times is highly time consuming. ANFIS modelling is done in about five seconds for up to three inputs, but can take up 30 min per turbine if seven inputs are used. Training NSET in this configuration requires usually 1-12 s with longer runtime for more inputs. MARS training is more expensive and can take more than a minute per turbine depending on the complexity of the model.

Evaluating the accuracy is feasible if the normal operation prediction is assessed. The error of prediction and actual measurement should be as small as possible and mean absolute errors, root mean squared errors, standard deviations or the coefficient of determination (R^2) can be used as metrics.

Table 4 compares the normal behaviour modelling accuracies based on the mean absolute error for the different modelling techniques, input selection cases a)-d) and the farms A-D. Due to the limitations in the service reports, the turbines in the study might be affected by further problems which could change the modelling performance, but the selection of the median value from all turbines should give a good indication of the accuracy. It can be seen, that NSET modelling is most accurate with mean absolute errors as low as 0.10 °C. This is due to the autoregressive nature of this technique. Using fewer inputs is better here, which can be explained by the stronger impact of the target signal in this case. All FSRC techniques are similarly accurate with slight advantages of ANN, ANFIS and MARS over LIN modelling. Using three instead of two inputs based on correlation usually improves the performance for FSRC techniques. Using only power and rotational speed to predict the drive train temperature is less accurate. Adding the ambient temperature as a third input improves the prediction in most cases. Comparing the different farms, prediction is most accurate in farm D, but similar low errors can be found in farms B and C. Prediction in farm A is less accurate, in particular in input case c). Possible reasons are manifold, but most likely the differences in the measurement setup and turbine operation of different manufacturers play a major role.

Table 3. Training time in seconds given as median values of all evaluated turbines and cases a) – d)

Technique	Farm A	Farm B	Farm C	Farm D
LIN	0.02 – 0.02	0.02 – 0.02	0.01 – 0.02	0.02 – 0.02
ANNs	1.73 – 3.02	1.41 – 2.00	1.58 – 1.99	1.70 – 2.93
ANFIS	4.42 – 5.70	4.40 – 5.42	4.39 – 5.43	4.40 – 5.48
NSET	1.27 – 1.94	2.40 – 6.64	1.90 – 5.09	3.00 – 11.52
MARS	23.81 – 71.82	5.44 – 25.42	1.37 – 18.77	0.78 – 11.00

Table 4. Accuracy in modelling normal behaviour given as median value of the mean absolute error (°C) of all evaluated turbines

Technique	Case	Farm A	Farm B	Farm C	Farm D
Used turbines		*102*	*8*	*18*	*6*
LIN	a)	2.24	1.22	0.98	1.07
	b)	2.03	0.96	0.91	0.85
	c)	11.41	1.62	1.39	1.53
	d)	3.07	1.29	1.06	1.45
ANNs	a)	2.27	1.12	0.87	0.89
	b)	2.00	0.89	0.86	0.82
	c)	9.47	1.58	1.39	1.19
	d)	3.21	1.27	1.41	1.49
ANFIS	a)	2.16	1.14	0.93	0.97
	b)	1.94	0.88	0.88	0.82
	c)	10.65	1.60	1.39	1.16
	d)	2.92	1.22	1.20	1.19
NSET	a)	0.23	0.33	0.38	0.91
	b)	0.25	0.27	0.35	1.09
	c)	0.18	0.10	0.10	0.13
	d)	0.25	0.77	1.53	0.52
MARS	a)	2.18	1.08	0.87	0.93
	b)	2.11	0.86	0.85	0.82
	c)	10.23	1.58	1.39	1.15
	d)	3.07	1.26	1.09	1.14

Accurate normal behaviour prediction does not necessarily imply good failure detection in terms of early and reliable alarms. Comparing the residual of modelled and measured temperatures before a known failure should give an insight in possible early warnings. However, displaying unfiltered residuals from a long period is not feasible due to the high number of samples per month and strong fluctuations.

In Fig. 2 – Fig. 6 the fortnightly moving averages of the residuals are given for 1.5 years' period before a gearbox replacement in farm B for LIN, ANN, ANFIS, NSET, MARS, respectively. Although most techniques and input selection cases show rising values in the last three months before the replacement, trends of a similar magnitude can be seen during the previous year. It is obvious that the smoothed residual is not an ideal indicator for failures and dedicated alarm generation techniques are required (which will be discussed below). However, ANN, NSET and MARS modelling with input case a) seem to give the most prominent increase directly before the replacement. Noticeably, case d) shows a trend from June to December which differs from the other cases in four of five modelling techniques. Future studies are required to better understand the impact of the input selection.

Modelling capabilities

These types of models are unlikely to be able to predict uncommon features in a signal. It has been observed that some abrupt increases in a bearing temperature in farm C cannot be modelled by any of the modelling techniques. These spikes occur only when the turbine power is rapidly increasing as shown in Fig. 7.

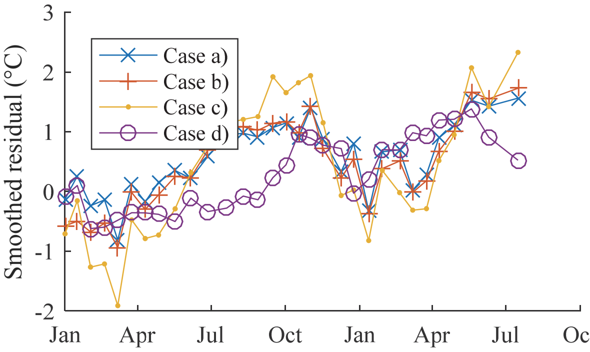

Fig. 2. Residuals from LIN modelling before a gearbox replacement

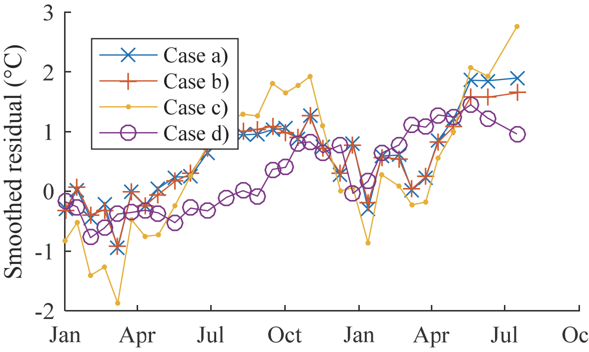

Fig. 3. Residuals from ANN modelling before a gearbox replacement

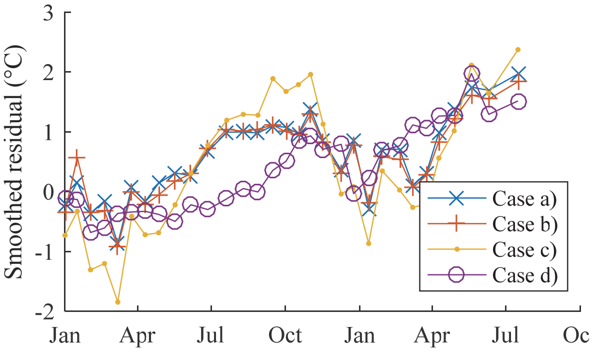

Fig. 4. Residuals from ANFIS modelling before a gearbox replacement

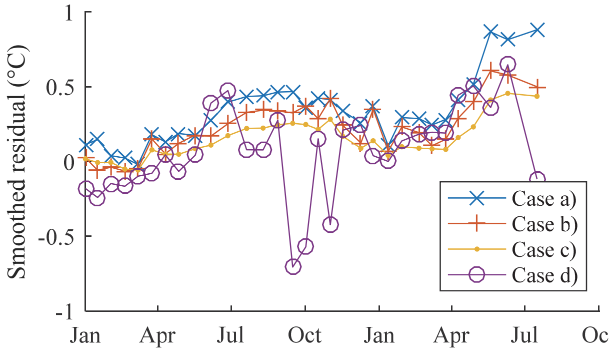

Fig. 5. Residuals from NSET modelling before a gearbox replacement

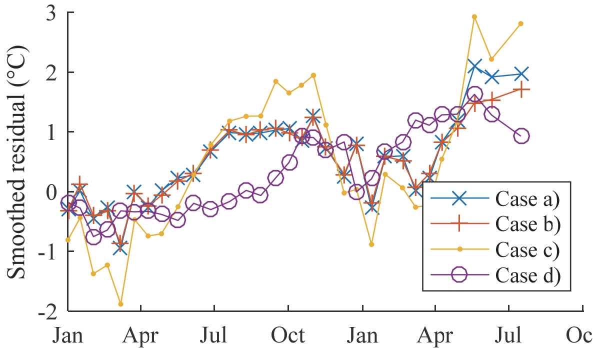

Fig. 6. Residuals from MARS modelling before a gearbox replacement

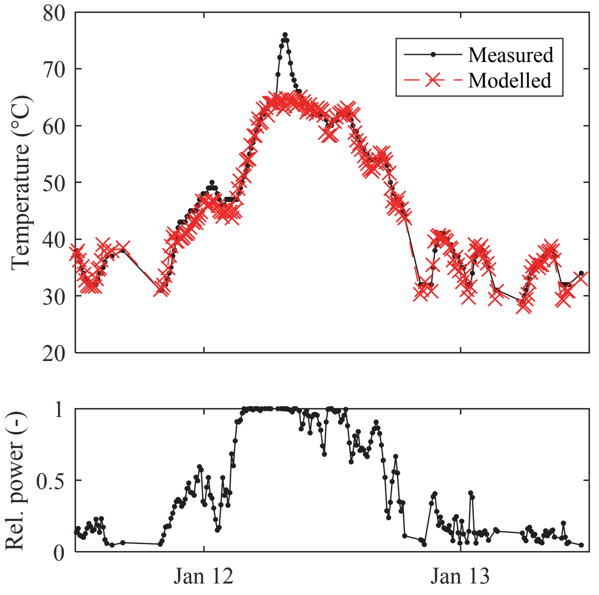

Fig. 7. Example of unpredicted spike in a bearing temperature when power is rapidly increasing (farm C, modelling with ANNs)

Alarm generation

The idea behind normal behaviour modelling is the use of the residual of measured minus modelled temperature to act as an indicator of potential failure. Different approaches have been proposed for generating alarms based on the residual time series.

Absolute threshold. The simplest way of generating alarms is by defining an absolute threshold for the residual. This can be by confidence bands (Garcia et al., 2006), with a defined threshold based on experience (Schlechtingen and Santos, 2010; Wilkinson et al., 2014) or a certain probability to occur derived from the error distribution in training, as e.g. less than 0.01 % (Schlechtingen et al., 2013). The reliability of the absolute threshold can be increased by using a daily average (Schlechtingen and Santos, 2010).

Mahalanobis distance. A Mahalanobis distance is a metric to condense the correlation of multiple variables and their distribution to a single number. Bangalore and Tjernberg (2015) proposed using a Mahalanobis distance of the residual and target referenced to the training distribution to detect anomalies. Alarms were raised if averages of three days were smaller than a distance with a probability of 1 % defined by a Weibull distribution fitted to the training results.

Exponentially weighted moving average control chart (EWMA). EWMA has been proposed to consider cumulating effects by Wang et al. (2016). Compared to the simple absolute threshold for the error, here a recursive statistic is built from the current error and the statistic in the previous time-step. A weighting of 0.2 for the current error and 0.8 for the previous statistic was used.

Abnormal level index (ALI). Sun et al. (2016) developed a numeric index to describe the abnormality of monitored signals. The index is calculated as a daily sum of penalties for residuals significantly bigger than the expected based on the training period. The penalty was defined as 5 and 3 for a penalty exceeding 97.5 and 75 % cumulative probability, respectively, or 1 else. After normalising, the index provided values between 0 and 1, with smaller values for less abnormality.

Discussion. Failure detection accuracy can be assessed in terms of the true positive and false positive alarms compared to the number of failures. Additionally, the advance time of detection before failure is a key measure. Comparing the failure detection capabilities is hindered by the above mentioned difficulties with the service reports. A thorough comparison of the modelling and the alarm generation techniques is out of the scope of this paper.

As an example, the detection of the above discussed gearbox failure in farm B modelled with ANNs in case a) is given in Fig. 8. All alarm generation techniques require a defined threshold for raising the alarm. The proposed probabilities of occurrence determined from the training data are not necessarily the optimal choice for new cases. A threshold defined by a probability of > 0.01 % for the residual leads to vanishing alarms in the investigated case studies. The limits are defined with a > 2 % probability of occurrence for the absolute daily threshold, > 1 % probability for the Mahalanobis distance, 6σ in the EWMA and the ALI as proposed. It can be seen that this selection of thresholds results in an increasing number of alarms in the two months before the replacement. However, a significant number of alarms occur far ahead of the replacement, in particular with the Mahalanobis distance. These alarms have to be considered as false alarms. The number of false alarms can be reduced by requiring several alarms in a row or in a specific time window as e.g. a week (Schlechtingen et al., 2013). By applying a limit of at least two days of alarms in one week and adapted thresholds the number of possibly false alarms can be reduced, as shown in Fig. 9. In this example, there is no significant difference between the first reliable detection of the different alarm generation techniques. However, the Mahalanobis distance and EWMA technique have a higher number of alarms after the first alarm than the absolute threshold and can be seen as more reliable accordingly. The fuzzy indicator provided by ALI shows a clear upward trend, but has two previous peaks which have to be assumed to be false alarms. A calibration of alarm thresholds with one turbine in the farm without replacements has been tested, but resulted in unsatisfactory results as many false alarms occurred.

The comparison of alarm generation techniques highlights that the adequate definition of thresholds has a higher impact than the differences in the detection approach. Future work has to address this problem in more detail.

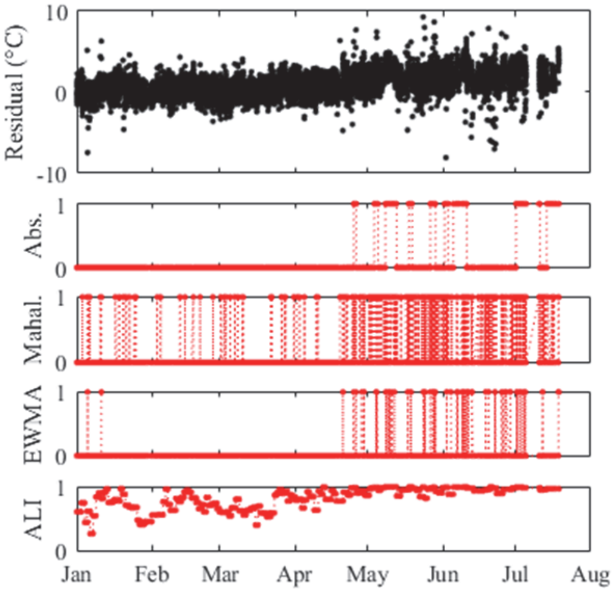

Fig. 8. Testing different alarm generation techniques (gearbox replacement at the end of time axis, ANNs modelling, case a), farm B)

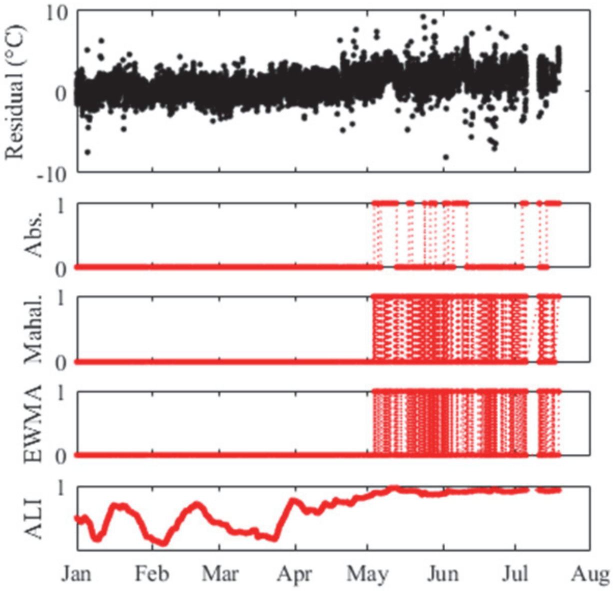

Fig. 9. Using a weekly filter for different alarm generation techniques (gearbox replacement at the end of time axis, ANNs modelling, case a), farm B)

FUTURE CHALLENGES IN ON-LINE MONITORING

Retrospective analyses of failures are of interest in an academic project of finding suitable tools for monitoring, but in industrial reality wind turbines are to be monitored on-line. Depending on the data management system this could be with new data every day, every ten minutes or even more frequently.

Most challenges which occur in retrospective analyses are also valid here, as e.g. SCADA data quality problems and model definitions. Problems with missing maintenance information are probably less severe in industrial on-line monitoring as a wind farm operator is aware of ongoing maintenance. However, insufficient or misleading maintenance reports occur in industrial practice too. The requirement for minimal computational effort for modelling will be even greater for on-line monitoring. Additionally, computational environments other than MATLAB are common in industry and will require adapted implementations. On-line monitoring will require adequate re-training of models after significant changes in the system or operation, as briefly discussed by Bangalore and Tjernberg (2013).

The main challenge of on-line monitoring is the required accuracy of monitoring in to allow decisions to be made about whether or not to send a maintenance team. A balance needs to be struck between providing early warnings whilst avoiding false alarms in order to minimise maintenance costs and maximise turbine availability. In the first months of operation of a new farm, there should be an iterative process of training and model evaluation until confidence in the results is achieved.

A combination of monitoring based on operational data and common vibration-based condition monitoring systems might be desirable to increase the reliability.

CONCLUSIONS

Based on analyses of case studies on four wind farms, the challenges in using operational data for wind turbine condition monitoring can be summarised as:

- Poor SCADA data documentation and quality,
- Insufficient maintenance documentation,
- The absence of best practice in selecting modelling techniques and settings,
- Isolated behavioural features which are difficult to model,
- The difficulty in defining sensible alarm thresholds which give sufficient notice of early genuine problems but minimise false alarms.

First findings indicate that ANN, ANFIS and MARS are similar accurate FSRC modelling techniques with the least computational effort for ANN. However, linear modelling is only slightly less accurate and possibly preferable due to its simplicity. NSET modelling is more accurate than all other techniques because of its autoregressive nature.

A brief comparison of smoothed residuals from all techniques before a gearbox replacement indicated that good modelling accuracy does not necessarily coincide with straightforward failure detection. Selecting inputs based on correlation or based on the physics seem to result in different residual trends, which are not fully understood yet.

If the different proposed alarm generation techniques are compared, no clear advantage of any approach is directly visible. A weekly filter of alarms is desirable to increase the certainty of results.

Future work will address the challenges in more detail and thoroughly evaluate the capabilities of failure detection with operational data.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642108 (Advanced Wind Energy Systems Operation and Maintenance Expertise, http://awesome-h2020.eu/).
We would like to thank the industrial partner in the AWESOME project, DNV GL, for sharing anonymised wind farm data.

REFERENCES

Bach-Andersen M, Rømer-Odgaard B, Winther O (2016). "Flexible non-linear predictive models for large-scale wind turbine diagnostics," *Wind Energy*, 17, 657–669.

Bangalore P, Tjernberg LB (2015). "An Artificial Neural Network Approach for Early Fault Detection of Gearbox Bearings," *IEEE Trans Smart Grid*, 6, 980–987.

Bangalore P, Tjernberg LB (2014). "Self Evolving Neural Network Based Algorithm for Fault Prognosis in Wind Turbines : A Case Study," *PMAPS*, 1–6.

Bangalore P, Tjernberg LB (2013). "An approach for self evolving neural network based algorithm for fault prognosis in wind turbine," In: 2013 IEEE Grenoble Conference. IEEE, pp 1–6.

Dienst S, Beseler J (2016). "Automatic Anomaly Detection in Offshore Wind SCADA Data," In: WindEurope Summit 2016. .

Feng Y, Qiu Y, Crabtree CJ, et al (2013). "Monitoring wind turbine gearboxes," *Wind Energy*, 16, 728–740.

Garcia MC, Sanz-Bobi MA, del Pico J (2006). "SIMAP: Intelligent System for Predictive Maintenance Application to the health condition monitoring of a windturbine gearbox," *Comput Ind*, 57, 552–568.

Guo P, Infield D, Yang X (2012). "Wind Turbine Generator Condition-Monitoring Using Temperature Trend Analysis," *Sustain Energy, IEEE Trans*, 3, 124–133.

Jekabsons G (2016). "ARESLab: Adaptive Regression Splines toolbox for Matlab/Octave," http://www.cs.rtu.lv/jekabsons/ (accessed 2016-03-24).

Schlechtingen M, Santos IF (2010). "Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection," *Mech Syst Signal Process*, 25, 1849–1875.

Schlechtingen M, Santos IF, Achiche S (2013). "Wind turbine condition monitoring based on SCADA data using normal behavior models. Part 1: System description," *Appl Soft Comput*, 13, 447–460.

Sun P, Li J, Wang C, Lei X (2016). "A generalized model for wind turbine anomaly identification based on SCADA data," *Appl Energy*, 168, 550–567.

Tan M, Zhang Z (2016). "Wind Turbine Modeling With Data-Driven Methods and Radially Uniform Designs," *IEEE Trans Ind Informatics*, 12, 1261–1269.

Tautz-Weinert J, Watson S (2016a). "Using SCADA Data for Wind Turbine Condition Monitoring - a Review," IET Renew Power Gener.

Tautz-Weinert J, Watson SJ (2016b). "Comparison of different modelling approaches of drive train temperature for the purposes of wind turbine failure detection," *J Phys Conf Ser*, 753, 72014.

Wang L, Zhang Z, Long H, et al (2016). "Wind Turbine Gearbox Failure Identification with Deep Neural Networks," *IEEE Trans Ind Informatics*, 3203, 1–1.

Wang Y, Infield D (2012). "Supervisory control and data acquisition data-based non-linear state estimation technique for wind turbine gearbox condition monitoring," *IET Renew Power Gener*, 7, 350–358.

Wiggelinkhuizen E, Verbruggen T, Braam H, et al (2008). "Assessment of Condition Monitoring Techniques for Offshore Wind Farms," *J Sol Energy Eng*, 130, 31004-1-31004–9.

Wilkinson M, Harman K, van Delft T, Darnell B (2014). "Comparison of methods for wind turbine condition monitoring with SCADA data," *IET Renew Power Gener*, 8, 390–397.

Yang W, Tavner PJ, Crabtree CJ, et al (2014). "Wind turbine condition monitoring: technical and commercial challenges," *Wind Energy*, 17, 673–693.

Zaher A, McArthur SDJ, Infield DG, Patel Y (2009). "Online wind turbine fault detection through automated SCADA data analysis," *Wind Energy*, 12, 574–593.