Bayesian state prediction of wind turbine bearing failure

Herp, Jürgen; Ramezani, Mohammad H.; Bach-Andersen, Martin; Pedersen, Niels Lovmand; Nadimi, Esmaeil S.

[Link back to DTU Orbit](#)

Bayesian state prediction of wind turbine bearing failure

Jürgen Herp [a, *], Mohammad H. Ramezani [a], Martin Bach-Andersen [b, c], Niels L. Pedersen [b], Esmaeil S. Nadimi [a]

[a] πSeG, ESRL, Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, 55 Campusvej, 5230, Odense, Denmark
[b] Siemens Wind Power A/S, Borupvej 16, 7330, Brande, Denmark
[c] Cognitive Systems, Technical University of Denmark, 2800, Kgs. Lyngby, Denmark

ARTICLE INFO

ABSTRACT

A statistical approach to abstract and predict turbine states in an online manner has been developed. Online inference is performed on temperature measurement residuals to predict the failure state Δn steps ahead of time. In this framework a case study is performed showing the ability to predict bearing failure 33 days, on average, ahead of time. The approach is based on the separability of the sufficient statistics and a hidden variable, namely the state length. The predictive probability is conditioned on the data available, as well as the state variables. It is shown that the predictive probability can be calculated by a model for the samples and a hazard function describing the probability for undergoing a state transition. This study is concerned with the prior training of the model, for which run-to-failure time series of bearing measurements are used. For the sample model prediction is conditioned on prior information and predict the next Δn samples from a feature space spanned by the prior samples. By assuming that the feature space can be described by a multivariate Gaussian distribution, the prediction is treated as a Gaussian process over the feature space.

1. Introduction

The continuous growth of wind energy generating sources, especially in harsh environments such as off-shore, has led to an increasing demand on more careful planning and control of operation and maintenance costs. This has made condition monitoring and fault diagnosis of wind turbines an even higher priority [1–5].

In this work failure of a turbine or it's components is defined as being the state of non-operation of aforesaid turbine or components. A fault on the other hand is associated with a defect, e.g. a crack in the bearings. As wind turbines are composed of different systems, it is intuitive that there are sub-systems which are more vulnerable to failure than others: the rotor system including the hub, has a more dominant failure rate than bearing and generator systems. However, the generator, hydraulic and gearboxes anecdotally are considered the *bête noires* of wind turbines, as, although not that frequent, the downtime caused by failure in these systems is substantial [3,6,7].

There are a wide variety of monitoring approaches available, ranging from acoustic analysis to visual inspection. These have shown potential in early fault detection, with prediction horizons ranging from seconds to months before a failure [3,8–16]. Amongst these are approaches that specific target bearing monitoring approaches [17,18]. Common causes for bearing failure are excessive load, fatigue, contamination, misalignments, overheating etc., latter will be addressed in the course of this paper. Additional common prediction, operation, and condition monitoring approaches are summarized in Kusiak et al. [19] and Márquez et al. [20].

As the proposed method in this work is a fault estimation (including prediction) approach, the comparison to other approaches in the field is essentially the comparison between the structure of set approaches. In general, fault estimation can be categorized into two groups: model based, and data-driven. In case of the first group, a physical model (such as Vidal et al. [21]) or at least an approximate state space model of the system (such as Gao et al. [22] and Liu et al. [23]) is necessary. Given initial information of the system and the consistency between the real and estimated variables, these methods have shown to be successful in providing robust fault estimation. The second group, on the other hand, is solely based on the recorded data, particularly suited when no

* Corresponding author.
    E-mail addresses: herp@mmmi.sdu.dk (J. Herp), ramezani@mmmi.sdu.dk (M.H. Ramezani), martin.bach-andersen@siemens.com (M. Bach-Andersen), niels_lovmand.pedersen@siemens.com (N.L. Pedersen), esi@mmmi.sdu.dk (E.S. Nadimi).

system informations are available. Although some of the data-driven methods employ a system model to generate residuals (as it is the case of this work), the employed models are unsupervised, such that no prior information of the system is used. The proposed approach in this study is data-driven and aims to predict baring fault based on the statistical features of residuals.

The work presented in this paper aims at achieving a high prediction horizon, but in contrast to other studies, emphasize the precision of the predictive model in order to provide a predictive horizon and the time of failure within a specified model accuracy. The prediction of a failure of a turbine or component is facilitated in predicting a specific fault. In the case study presented in this work, this is equal to the prediction of the remaining lifetime until failure of the turbine, as run-to-failure time series are considered. This is achieved by abstracting turbine states in a Bayesian framework [24]. The contribution of this work is to extend the work of Herp et al. [24] by including a prediction horizon based on modelling samples as part of a Gaussian process, associated with bearing failure. Further, training a predictive model by updating the hyper-parameters, on run-to-failure time series, is added in this work as well. Investigating faults beyond bearing failure, and especially unsupervised fault identification and isolation, such as benchmark models by Dey et al. and Odgaard et al. [15,25,26] will not be addressed in this study. This challenge is left for consideration in future work.

The paper is organized as follows: In Section 2 the problem at hand is formulated and the terminology is introduced. Further, a brief description of the data is made available. Section 3 introduces the state transition approach and its extension considering Gaussian processes. Following the theory of Section 3 the state prediction is described in Section 4. The proposed approach is put to use in a case study on bearing failures in Section 5. Finally, Section 6 concludes the outcome of this paper.

2. Formalizing the motivation

Recent work has explored the idea of abstracting states from wind turbine measurement residuals [24,27]. Under the assumption that wind turbine states can be characterized by changes in the mean and standard deviation of a time series, a recursive online algorithm was proposed. Motivated by this work, the question *when will a given state occur?* will be addressed.

Although this study is concerned with run-to-failure residuals of bearing temperature, the problem will be formulated in a general way to highlight the universality of the algorithm. Consider E_k to be any event which is linked to the operation of a wind turbine at time i, dependent on a hidden variable $s^{(i)}$, representing the current state of the turbine as well as the data collected up to time i. The question raised is, what is the probability of detecting E_k or a set of events $\{E_k\}$, Δn samples before its occurrence, $\mathbb{P}(\Delta n | \{E_k\}, s^{(i)}, \mathbf{x}_{[1,i]})$, conditioned on the prior samples and current state variable? Here, $\mathbf{x}_{[a,b]}$ defines the discrete set of successive observations from time a to b. Assuming the separability of the process $\{x_1, ..., x_N\}$ into $\mathscr{S} \leq N$ states, $s^{(i)}$ will characterise the length of the current state with samples $\mathbf{x}_{[S_{\nearrow}, i]}$. s_m referrers to the hidden variable of a wind turbine, defined by the separability of the process and will also be used to label the states. The transition between those states are referred to as state transitions.

Consider $\{E_{k''}\}$ as the set of events associated with bearing temperature failure, the scope of this work, is to show that $\mathbb{P}(\Delta n | \{E_{k''}\}, s^{(i)}, \mathbf{x}_{1,i})$ can be calculated by extending the work of Herp et al. [24] to a wider range of models. The model extension will focus on the principal of Gaussian processes (Rasmussen et al. [28]).

2.1. Run-to-failure bearing time series

The data at hand can be divided into two different types: operational data, in the form of preprocessed Supervised Control And Data Acquisition (SCADA) data, and status data, referred to as event data as mentioned earlier.

2.1.1. Residuals of bearing temperature

Three time series of bearing temperature residuals, ΔT, are considered for this study and presented in Fig. 1. Summarizing from the work of Bach-Andersen et al. [18], the underlying model generating the residuals is not meant to resemble a physical model of the wind turbine's thermal energy flow, hence an artificial neural network (ANN) is chosen for modelling the temperatures of the main bearing component. The input for the model is chosen based on first principle consideration of the energy diffusion in a wind turbine, subjected to the bearing assembly. The gear box is considered the primary thermal source, the diffusion follows through the main shaft and bearing assembly, before it is dissipated to the outside of the hub. The model inputs are SCADA data sampled in 10min intervals, containing active power, generator speed, gear box oil temperature, ambient temperature, and nacelle temperature. The model output is re-sampled to hourly resolution. Furthermore, the ANN model is trained on assumable fault-free data from the first year of operation for each turbine.

Besides the common behaviour explored in Fig. 1, it is essential that the time series under considerations show common statistics, in order to utilize machine learning concepts of training hyper-parameters for state prediction. Plotting the quantiles of the time series against the quantiles of a normal distribution, Fig. 2, shows that the main part of all three time series follow a normal distribution with zero mean and unit variance. Further, Fig. 2 indicates that the buildup in temperature follows a similar distribution in all
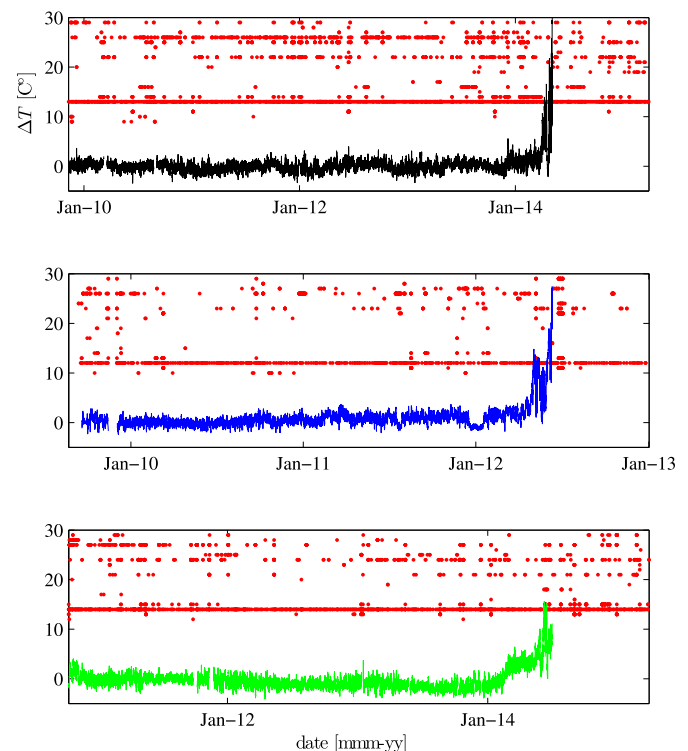

Fig. 1. Run-to-failure time series of bearing temperature for Turbine 1 (top), Turbine 2 (mid) and Turbine 3 (lower), including recorded events (●). Each row represents a different type of event group.
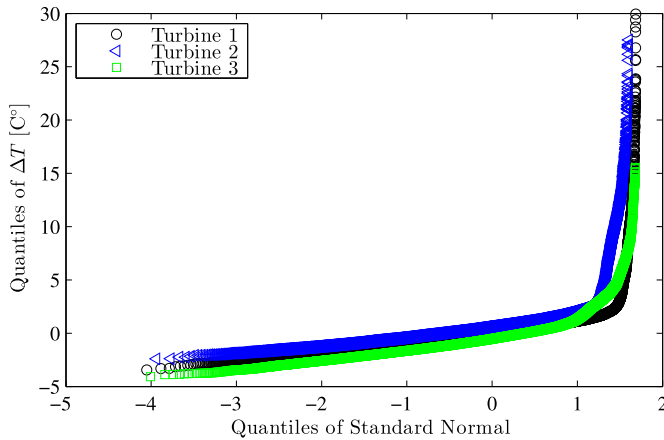
Fig. 2. Quantiles of the bearing residuals for the turbines under consideration compared to the quantiles of a standard normal distribution.

time series. As such, the assumption is made that we can train on one or more time series and apply the obtained model on similar time series without adjusting the model.

2.1.2. Turbine event data

Event data consists of the recorded events experienced by a wind turbine, e.g. failure, warning, status etc., and the associated timestamps. These events can be mapped binary into any time frame matching the SCADA system as seen in Table 1. In order to give the reader an overview of how many events occur, Fig. 1 indicates, by red markers, all events experienced by the wind turbines, ordered in groups. These events range from scheduled lubrications, over untwisting cables, to bearing over-temperature. Remark, on average around 1000 occurrences per year of different events are under consideration in this study.

As failures of wind turbines develop over time, it is reasonable to assume that event patterns might be abstracted in order to justify temporal correlations amongst a turbine's states. Considering a chain of successive and/or simultaneous events, the dependency between two sets of events is referred to as event patterns and

formulated as

$$
\left.\begin{array}{ll}
\mathscr{E}_l \Rightarrow \mathscr{E}_{l'} & l \neq l' \\
\mathscr{E}_l, \mathscr{E}_{l'} \subseteq \mathscr{E} \\
\mathscr{E}_l \cap \mathscr{E}_{l'} \neq \varnothing
\end{array}\right\}, \tag{1}
$$

with support (relative number of a pattern), and confidence

$$
\mathrm{conf}(\mathscr{E}_l \Rightarrow \mathscr{E}_{l'}) = \frac{\mathrm{supp}(\mathscr{E}_l \cup \mathscr{E}_{l'})}{\mathrm{supp}(\mathscr{E}_l)}. \tag{2}
$$

By this approach, normal and critical event patterns of bearing failure can be abstracted from the turbines under consideration. For instance, with high confidence, *Bearing Over-Temperature* warnings can lead to *Bearing Over-Temperature* failures.

3. State transition

The state transition algorithm under consideration follows the work of Herp et al.[24, 27] where the joint probability over the current state can be estimated recursively by

$$
\mathbb{P}\left(s^{(i)}, \mathbf{x}_{[1,i]}\right) = \sum_{s^{(i-1)}} \underbrace{\mathbb{P}\left(s^{(i)}\middle| s^{(i-1)}, \mathbf{x}_{[1,i]}\right)}_{\text{conditinal prior}} \times \underbrace{\mathbb{P}\left(x_i\middle| s^{(i-1)}, \mathbf{x}_{[1,i-1]}\right)}_{\text{sample model}}
$$
$$
\times \mathbb{P}\left(s^{(i-1)}, \mathbf{x}_{[1,i-1]}\right). \tag{3}
$$

Eq. (3) is referred to as the Dependent State Transition Model (DSTM). The DSTM can be easily generalized to higher dimensions by portraying x_i as a vector $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}_{[1,i]}$ as a matrix of dimensions $i \times d$. States are then abstracted by considering the maximum likelihood of $\mathbb{P}(s^{(i)}, \mathbf{x}_{[1,i]})$. Both the conditional prior and sample model are implicitly depending on known hyperparameters $\boldsymbol{\beta} = (\beta_c, \beta_m)$, thus whenever one of them is considered it is referred to as $\mathbb{P}(s^{(i)}|s^{(i-1)}, \mathbf{x}_{[1,i]}) \equiv \mathbb{P}(s^{(i)}|s^{(i-1)}, \mathbf{x}_{[1,i]}, \beta_c)$ or $\mathbb{P}(x_i|s^{(i-1)}, \mathbf{x}_{[1,i-1]}) \equiv \mathbb{P}(x_i|s^{(i-1)}, \mathbf{x}_{[1,i-1]}, \beta_m)$.

The probability of undergoing a state transition is based on the

Table 1
Binary Event Mapping: Mapping from recorded event data into a binary array for the evaluation of dependency between events.

Event	Start Time	Stop Time
E_1	Oct. 11, 2014, 12:42:12	Oct. 12, 2014, 16:02:22
E_{21}	Oct. 12, 2014, 09:59:01	Oct. 13, 2014, 10:12:43
E_1	Oct. 12, 2014, 15:39:06	Oct. 13, 2014, 10:22:43
$\vdots$	$\vdots$	$\vdots$
E_3	Oct. 19, 2014, 02:22:00	Oct. 19, 2014, 13:00:00
E_{12}	Oct. 19, 2014, 03:01:52	Oct. 23, 2014, 13:00:00

$$\Downarrow$$

sample	E_1	E_2	$\cdots$	E_{k-1}	E_k	E_{k+1}	$\cdots$	E_K
$i-2$	0	0	$\cdots$	0	1	0	$\cdots$	1
$i-1$	0	0	$\cdots$	0	1	0	$\cdots$	1
i	0	0	$\cdots$	0	1	1	$\cdots$	0
$i+1$	0	0	$\cdots$	0	0	1	$\cdots$	0
$i+2$	1	0	$\cdots$	0	0	1	$\cdots$	1

The column header "Events" spans E_1 through E_K.

assumption of dependencies between states. Referring to Herp et al. [24] for the calculation of the DSTM the dependency on $\mathbf{x}_{[1,i]}$ requires taking the previous state information into account. I.e. utilizing the knowledge of the number of previous state transitions $\mathscr{S}$, and their location in the sequence of states: $S_0 = 0 < S_1 < S_2 < \ldots < S_{\mathscr{S}} < i$. The estimation of the conditional prior in Eq. (3), will be based on the temporal distance between the occurrence of states, $\Delta S \equiv S_m - S_{m-1}$. In other words, $f(\Delta S) = \mathbb{P}(S_m - S_{m-1} = \Delta S) \sim \{\mathbb{N}_1\}$, where $1 \leq m < i$. As the event dependencies do not contain temporal information, the assumption that ΔS can be treated as an independent variable, with the joint probability distribution, Eq (4), is implied.

$$\mathbb{P}(S_1, \ldots, S_{\mathscr{S}}) = \left(\prod_{m=1}^{S} f(\Delta S) \right) \left(1 - F\left(s^{(i)}\right) \right), \tag{4}$$

here F denotes the cumulative distribution of f. When considering the ith sample, the dataset is divided into $\mathscr{S} + 1$ sates, with the mth state containing the observations $\mathbf{x}_{[S_m+1, S_{m+1}]}$, up to $m = \mathscr{S}$. For each state there then exists a model specified by its statistical measures. For all m the probability distribution over s_m will depend on the location of the previous state transition and its length, characterized by (S_{m-1}, S_m), and the statistical information in s_{m-1}. Let $\mathbb{P}(s^{(i)} | s^{(i-1)}, \mathbf{x}_{[1,i]})$ be equal to H, the probability when undergoing a state transition, $1 - H$ when staying in the same state, and 0 otherwise. The probability when undergoing a state transition will be proportional to $\mathbb{P}(s_m | s_{m-1}, S_m, S_{m-1})$, and the statistical information in S_{m-1}. Remark, in this cases common parameters across the states are possible. In order to initialize the first state, s_0, the prior for this state needs to be picked beforehand, since no data are available yet. Finally, the problem is reduced to selecting the proper set of conjugated priors (as discussed in a wide range of textbooks such as Gelman et al. [29]) for s_m and S_m, that best resembles the data set.

The sample model in Eq. (3) describes the statistical measures of the different states. The original state transition approach considered changes in mean and variance to distinguish between states. A Student's t-distribution can be used for that purpose. Let $\text{St}_\nu(\mu, \sigma)$ denote the Student's t-distribution with ν degrees of freedom, mean μ, and variance σ, then

$$\mathbb{P}\left(x_i \Big| s^{(i-1)}, \mathbf{x}_{[1,i-1]}\right) = \text{St}_{2\alpha}\left(\mu_i, \frac{\gamma_i}{\alpha_i} \frac{\kappa_i + 1}{\kappa_i}\right), \tag{5}$$

where

$$\mu_i = \frac{\kappa_0 \mu_0 + \mathbb{E}\left[\mathbf{x}_{[1,i]}\right]}{\kappa_0 + i}, \tag{6a}$$

$$\kappa_i = \kappa_0 + i, \tag{6b}$$

$$\alpha_i = \alpha_0 + i, \tag{6c}$$

$$\gamma_i = \gamma_0 + \frac{1}{2}\xi_i + \frac{\kappa_0 i\left(\mathbb{E}\left[\mathbf{x}_{[1,i]}\right] - \mu_0\right)^2}{2(\kappa_0 + i)}, \tag{6d}$$

with $\mu_0, \kappa_0, \alpha_0, \gamma_0$ being the previous statistics, and sum of standardised squared error

$$\xi_i = \sum_{l=1}^{i} (x_l - \mathbb{E}[x])^2. \tag{7}$$

As seen in Section 2.1; time series under consideration might not show characteristics that can be described by simple changes in their mean or variance. In contrast to the approach proposed by Herp et al. [24], this work considers more elaborate models by the means of Gaussian processes as described in the next section.

3.1. Gaussian process - extensions to the existing model

Gaussian processes have been proven to be useful in machine learning, a profound description is available in the textbook of Rasmussen et al. [28]. By definition, a Gaussian process is a collection of random variables of finite number which have a Gaussian distribution. The Gaussian process will thus be fully specified by a mean $\mathbb{E}[\cdot] \in \mathbb{R}^d \to \mathbb{R}$ and covariance $\sigma(\cdot, \cdot) \in \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ function with known hyper-parameters λ and input space dimension d. In this study, the sample model in Eq. (3) is described by a Gaussian process. Defining the input space as the space of training sample data, and the feature space as the space of sample data of the system under study, these spaces might have different dimensions. Considering $\omega \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as weight to the input, $\phi(i)$ is introduced as the function mapping from the d-dimensional input space into a finite sized feature space of dimension D:

$$f(i) = \phi(i)\omega. \tag{8}$$

Now $f(i)$ can be specified by the mean and the covariance of the feature space:

$$\mathbf{x}_i = f(i) + \varepsilon_i, \tag{9a}$$

$$f_i \equiv f(i) \sim \mathscr{GP}\left(\mu_f, \sigma_f\right) \equiv \mathcal{N}\left(\mu_f, \sigma_f\right), \ \varepsilon_i \sim \mathcal{N}\left(0, \mathbb{I}\sigma_\varepsilon^2\right), \tag{9b}$$

where

$$\mathbb{E}[f_i] = \phi(i)^\top \mathbb{E}[\omega] =: \mu_f, \tag{10}$$

$$\mathbb{E}\left[f_i, f_j\right] = \phi(i)^\top \mathbb{E}\left[\omega\omega^\top\right]\phi(j) = \phi(i)^\top \boldsymbol{\Sigma}\phi(j) =: \sigma_f. \tag{11}$$

By construction of Eq. (3) the sample model is only dependent on the current state, i.e. $\mathbf{x}_{[S_{\mathscr{S}}+1, i-1]}$. In the light of Gaussian process, the covariance between the predicted sample, i.e. $\mathbf{x}_i$, and the current state, the prior can be written as:

$$\begin{bmatrix} \mathbf{x}_{[S_{\mathscr{S}}+1, i-1]} \\ \mathbf{x}_i \end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix} \sigma\left(\mathbf{x}_{[S_{\mathscr{S}}+1, i-1]}, \mathbf{x}_{[S_{\mathscr{S}}+1, i-1]}\right) + \sigma_\varepsilon^2 \mathbb{I} & \sigma\left(\mathbf{x}_{[S_{\mathscr{S}}+1, i-1]}, \mathbf{x}_i\right) \\ \sigma\left(\mathbf{x}_{[S_{\mathscr{S}}+1, i-1]}, \mathbf{x}_i\right)^\top & \sigma(\mathbf{x}_i, \mathbf{x}_i) \end{bmatrix}\right). \tag{12}$$

From Eq. (12) the sample model in Eq. (3) (in earlier work described by a Student's t-distribution) gives rise to the following one step predictive distribution, by simple matrix manipulation [30]:

$$\mathbb{P}\left(\mathbf{x}_i \big| \mathbf{x}_{[S_{\nearrow}+1,i-1]}\right) = \mathcal{N}(\mu, \text{cov}), \tag{13}$$

where

$$\mu = \sigma\left(\mathbf{x}_{[S_{\nearrow}+1,i-1]}, \mathbf{x}_i\right)^\top \cdot \left(\sigma\left(\mathbf{x}_{[S_{\nearrow}+1,i-1]}, \mathbf{x}_{[S_{\nearrow}+1,i-1]}\right)\right.$$
$$\left. + \sigma_\varepsilon^2 \mathbb{I}\right)^{-1} \mathbf{x}_{[S_{\nearrow}+1,i-1]}, \tag{14}$$

$$\text{cov} = \sigma(\mathbf{x}_i, \mathbf{x}_i) - \sigma\left(\mathbf{x}_{[S_{\nearrow}+1,i-1]}, \mathbf{x}_i\right)^\top$$
$$\cdot \left(\sigma\left(\mathbf{x}_{[S_{\nearrow}+1,i-1]}, \mathbf{x}_{[S_{\nearrow}+1,i-1]}\right) + \sigma_\varepsilon^2 \mathbb{I}\right)^{-1} \cdot \sigma\left(\mathbf{x}_{[S_{\nearrow}+1,i-1]}, \mathbf{x}_i\right). \tag{15}$$

Consider again the signal plus noise model of the form $\mathbf{x}_i = f(i) + \varepsilon_i$, remark that the additive noise is assumed to be independent, identical distributed with variance σ_ε^2. The prior on the noise part can be expressed as $\sigma(\mathbf{x}_i, \mathbf{x}_j) + \sigma_\varepsilon^2 \mathbb{I}$. The joint predictive distribution for the samples of some prior input $\mathbf{x}_{[S_{\nearrow}+1,i-1]}$ and predictive output $\mathbf{x}_i$ is specified by the prior.

It is assumed that the hyper-parameter of the sample model, namely β_m, is known and fixed. In the cases of multiple hyper-parameters γ for each state transition, these need to be marginalized out

$$\mathbb{P}\left(\mathbf{x}_i \big| \mathbf{x}_{[S_{\nearrow}+1,i-1]}, \beta_m\right) \propto \int \mathbb{P}\left(\mathbf{x}_i \big| \mathbf{x}_{[S_{\nearrow}+1,i-1]}, \gamma\right) \mathbb{P}(\mathbf{x}_i|\gamma) \mathbb{P}(\gamma|\beta_m) d\gamma, \tag{16}$$

with marginal log-likelihood

described in Section 3, into prediction of state transition conditioned on a single event is based on two principles: (i) the common characteristics across the time series such that one or more turbines can be utilized for model building purposes, i.e. to learn of the underlying hyper-parameters and (ii) that the future samples, $\mathbf{x}_{i+\Delta n}$ can be extrapolated by the Gaussian process as described by Eq. (12).

(i) The hyper-parameters are obtained by considering one or more training time series and running the Gaussian process built around Eq. (3) such that $\arg \max_\beta \{L(\mathbb{P}(S^{(i)}|\mathbf{x}_{[1,N]}))\}$, where $\mathbf{x}_{[1,N]} = [\mathbf{x}_{[1,i]}, \mathbf{x}_{[i+1,i+\Delta N]}]$. In order to assure model robustness and prevent over-fitting, the training process can be done on more than one time series. In that case, multiple time series are taken as an input to retrieve the hyper-parameters. The training time series are aligned at the time of failure and treated as a multivariate input for each $\mathbf{x}_i$, i.e. following the approach outlined in Section 3 with $d \leq 2$. Compared to the previous work on the state abstraction of wind turbines [24,27], where training was only performed on a subset of a time series, training on a full time series now enables the prediction of specific states in other time series. This is due to the trained model's awareness of the possible states existence in the future.

(ii) As soon as the hyper-parameters are specified, time series can be predicted Δn samples into the future. Intuitively speaking, the prediction can be viewed as the extrapolation of the samples $\mathbf{x}_{[1,i-1]}$ by the functions $f(\cdot)$ selected from the feature space. For any $\mathbf{x}_{[1,i-1]}$ Eq. (12) returns the estimated $\mathbf{x}_i$. Iteratively applying Eq. (12) leads to $\mathbf{x}_{[1,i]}, \mathbf{x}_{[1,i+1]}, \ldots, \mathbf{x}_{[1,i+\Delta n]}$. The hyper-parameters for the sample model, β_m, or equivalently $\{\gamma\}$, dictate the properties of the hidden state variable s_m, i.e. ΔS, the length of a state, and its location (S_{m-1}, S_m).

Fig. 3 illustrates the prediction of $\mathbf{x}_i$ for simplified and idealized

$$\log \mathbb{P}\left(\mathbf{x}_{[S_{\nearrow}+1,i-1]} \big| \gamma\right) = -\frac{1}{2}\left[\mathbf{x}_{[S_{\nearrow}+1,i-1]}^\top \cdot \left(\sigma\left(\mathbf{x}_{[S_{\nearrow}+1,i-1]}, \mathbf{x}_{[S_{\nearrow}+1,i-1]}\right) + \sigma_\varepsilon^2 \mathbb{I}\right)^{-1} \mathbf{x}_{[S_{\nearrow}+1,i-1]} + \log\left(\left|\sigma\left(\mathbf{x}_{[S_{\nearrow}+1,i-1]}, \mathbf{x}_{[S_{\nearrow}+1,i-1]}\right) + \sigma_\varepsilon^2 \mathbb{I}\right|\right)\right.$$
$$\left. + (S_{\nearrow} + 1)\log(2\pi)\right]. \tag{17}$$

For this study $d = 1$, unless stated otherwise.

Generating samples using the Gaussian process based sample model requires the selection of specific structures for μ and cov, and estimating these structures by obtaining the hyper-parameters on a training data set. While μ is assumed to be fixed and zero, and the prior on the covariance function, cov, is chosen to be a *radial basic function* [31] of the form

$$\text{cov} = \sigma(\mathbf{x}_i, \mathbf{x}_j) = \sigma_\varepsilon^2 \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_j - \mathbf{x}_i)\right\}, \tag{18}$$

generating infinitely differentiable functions.

4. State prediction

Extending the Gaussian process detection of state transitions, as

simulated bearing temperature residuals with additive noise. The predictions shown in panels a) - d) are based on Eq. (12), generating the prediction, shown in gray-scale (━), based on the previous samples. Early prior to the buildup in temperature (panel a)), the prediction indicates no significant change in the upcoming residuals. As time progresses the algorithm picks up the buildup in temperature, and predicts the rise of the temperature as it looks further into the future (panel b)). Panel c) shows how the precision decreases and the predictive distribution hardens around the predictive values. Finally, as seen in panel d), the turbine undergoes downtime after a fault occurred, and the temperature falls to normal levels again.

Being able to predict future samples, the prediction horizon, accuracy, and precision can be addressed, under the assumption of monitoring a specific event $E_{k''}$ or set of events $\{E_{k''}\}$. As β_m, or equivalently $\{\gamma\}$, specify the state transitions, it is possible to evaluate $\mathbb{P}(\Delta n|\{E_{k''}\}, s^{(i)}, \mathbf{x}_{[1,i]})$. Fig. 4, illustrates that the idealized
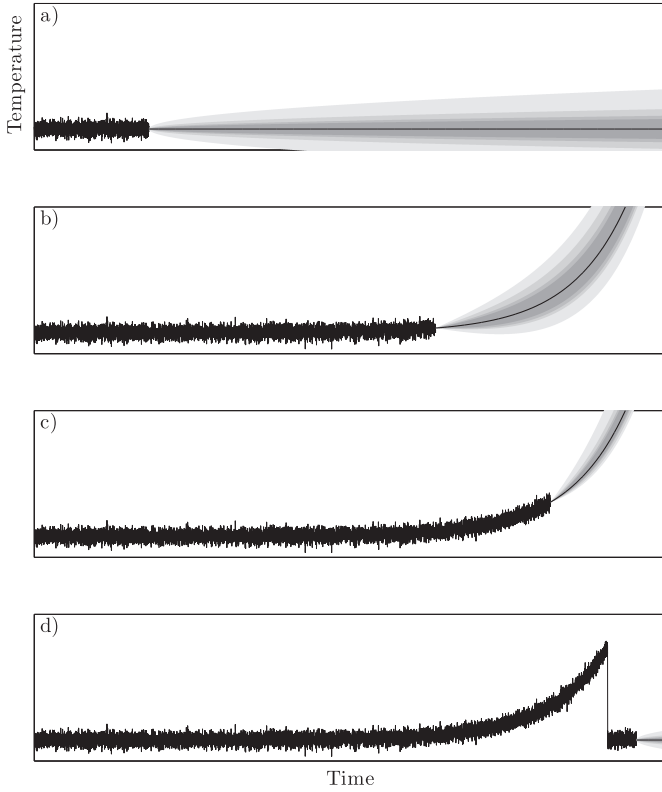
Fig. 3. Concept of state, the Gaussian process predictive model and state prediction. Snapshots of the predictive model, forecasting samples according to the trained Gaussian process, with variance indicated in gray-scale (━).

distribution $\mathbb{P}(\Delta n | \{E_{k''}\}, s^{(i)}, \mathbf{x}_{[1,i]})$ will have an empirical probability density function (pdf(Δn)) and cumulative distribution function (cdf(Δn)). The maximum likelihood of $\mathbb{P}(\Delta n | \{E_{k''}\}, s^{(i)}, \mathbf{x}_{[1,i]})$ defines the predicted time, t_n, of $E_{k''}$. The prediction horizon is defined as $t_n - t_i$, the distance between the current time t_i and the most likely time of the event t_n dictated by pdf(Δn). The accuracy is defined as

$$\text{ACC} = \left| t_{E_{k''}} - t_n \right|, \tag{19}$$

the absolute distance between the real time, $t_{E_{k''}}$, and the predicted
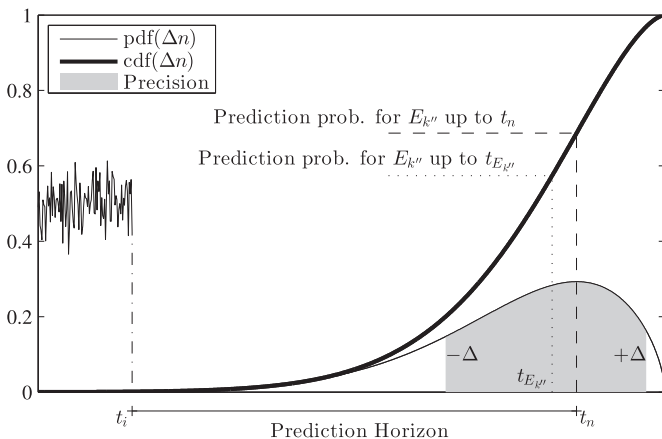

Fig. 4. Idealized Δn-step prediction of an event. Illustrated are the probability density function and cumulative distribution function from a Gaussian process prediction a specific event.
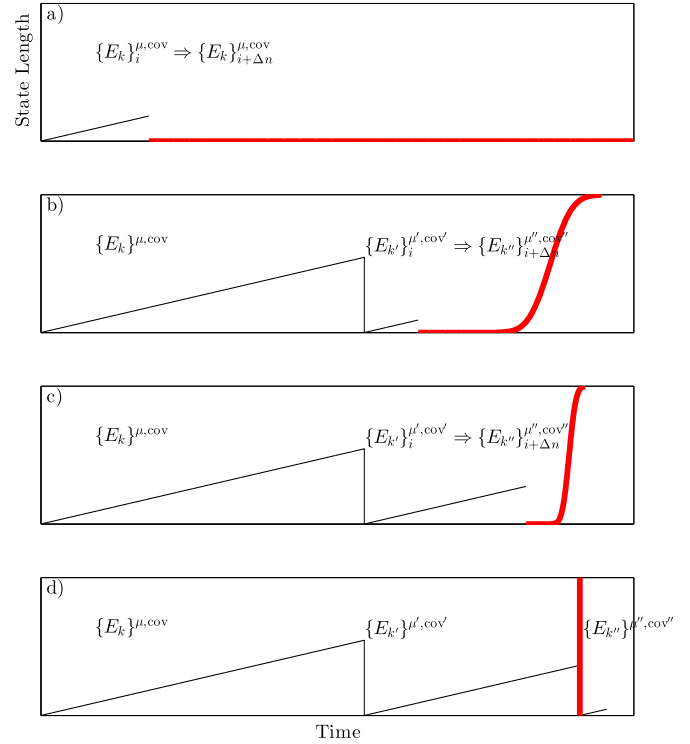

Fig. 5. Concept of state, the Gaussian process predictive model and state prediction. Predicted events based on the event pattern analysis, and cumulative distribution for observing specific events $\{E_{k'}\}$. Further, the diagonal lines indicate the abstracted states.

time for the considered event. The precision of predicting t_{E_k} is defined as the symmetric integral over $\mathbb{P}(\Delta n | \{E_{k''}\}, s^{(i)}, \mathbf{x}_{[1,i]})$:

$$\text{PRC} = \int_{t_{E_{k''}} - \Delta}^{t_{E_{k''}} + \Delta} \text{pdf}(\Delta n)dt \rightarrow \sum_{t_{E_{k''}} - \Delta}^{t_{E_{k''}} + \Delta} \mathbb{P}\left(\Delta n \middle| E_{k''}, s^{(i)}, \mathbf{x}_{[1,i]}\right) \tag{20}$$

where Δ indicates the desired precision interval around $t_{E_{k''}}$, as illustrated by the shaded area in Fig. 4. The probability density function, pdf(Δn), in Fig. 4 is best explained intuitively by considering a hard temperature threshold. In the case where the threshold cuts through the shaded area in Fig. 3 it will give raise to a distinction as shown in Fig. 4.

$\{\cdot\}_t^{\mu,\text{cov}}$ is introduced as the short hand notation for the set of events associated with the mean and variance of the predictive distribution $\mathcal{N}(\mu, \text{cov})$, as defined by Eqs. (13) - (15). Assuming the events under consideration, $\{E_{k''}\}$, are fixed, states can be predicted under the premises of Eq. (1), such that $\{\cdot\}_t^{\mu,\text{cov}} \Rightarrow \{\cdot\}_{t+\Delta n}^{\mu',\text{cov}'}$ expresses the patterns of Eq. (1). Continuing the discussion of the idealized bearing temperature residuals prediction of Figs. 3 and 5 shows the abstracted state events and the prediction, conditioned on the monitoring of $\{E_{k''}\}$. Each row in Fig. 5 is described separately and correspond to the panels a)-d) in Fig. 3:

a) As the predicted samples are a linear extrapolation of the past interval, the state detection algorithm does not identify any state transition at any time in the future. This is expressed in the close to zero cdf(Δn). Further, as μ and cov do not change, the predicted events Δn samples ahead of i identical to the set of the current events, i.e. $\{E_k\}_i^{\mu,\text{cov}} \Rightarrow \{E_k\}_{i+\Delta n}^{\mu,\text{cov}}$.

b) As the predicted sample changes, a state transition is detected. The new prediction generates samples that are associated with a temperature increase. As such, the cumulative distribution of $\mathbb{P}(\Delta n | \{E'_k\}, s^{(i)}, \mathbf{x}_{[1,i]})$ starts building up. The change of the Gaussian process, through μ and cov, identifies the current state as a *pre-fault* $\{E_k\}_i^{\mu'',\mathrm{cov}}$, i.e. a state prior to a failure, predicting the next state to contain events associated with the turbine failure, $\{E_k\}_i^{\mu'',\mathrm{cov}''}$. The cumulative distribution indicates when the predictive algorithm expects the next state transition, t_n, associated with $\{E_k\}_i^{\mu'',\mathrm{cov}''}$. Notice that the previous state does not contribute to the prediction anymore and is completely specified by μ, cov and $\{E_k\}$.

c) This row describes the same situation as row b), including the same event patterns. However, time is progressed and more samples are available. In the snapshot presented, the distribution around the predicted values has hardened, leading to a narrower pdf(Δn) and a more local cdf(Δn). This is equal to the state prediction is more certain of predicting a state transition at a specific location in time. Remark, for the state prediction approach it cannot be shown that the prediction will converge fully towards the true failure, upon increasing sample size and decreasing distance to the failure. This is due to the highly complex data driven nature of the hyper-parameters involved across the states. The next Section will provide the implementation of the proposed approach on the earlier mentioned residuals.

d) Finally the prediction of a state, subjected to $\{E_k\}_i^{\mu'',\mathrm{cov}''}$, has converged toward a point in time and a new state transition is detected. The cumulative distribution at that point becomes a step function.

Summarizing the panels a) - d) of Fig. 5, they show the convergences of the state prediction towards the failure under consideration. Remark, Figs. 3 and 5 are idealized cases for illustrational purposes only. The computational effort for estimating the probability of each possible state is in the order of $\mathcal{O}(n^2)$, where n is the length of the time series.

5. Bearing failure - A case study

The time series described in Section 2.1 will now be under investigation, by applying the approach laid out in Sections 3 and 4. Further, following the notation of Section 4, $\{E_{k''}\}$ denotes the set of events which are associated with *Bearing Over-Temperature* failure, and $\{\cdot\}_t^{\mu,\mathrm{cov}} \Rightarrow \{\cdot\}_{t+\Delta n}^{\mu',\mathrm{cov}'}$ are the patterns extracted under the premises of Eq. (1), exceeding the heuristic confidence threshold of ≥ 0.75. When transitioning into a state associated with bearing failure, i.e. $\{E_{k'}\}_t^{\mu',\mathrm{cov}'} \Rightarrow \{E_{k''}\}_{t+\Delta n}^{\mu'',\mathrm{cov}''}$, the averaged confidence over all turbines is 0.79. Before evaluating the times series, the state detection and prediction approach is built on the characteristics of the remaining times series, in accordance with the assumed similarity presented in Figs. 1 and 2. For the remainder of this section one time series is under consideration at a time, leaving the other two time series for the training objective. By permutation of the training time series, the training input can be of dimension $d = 1, 2$. For example, evaluating the temperature residuals of Turbine 1, $d = 1$ for training the model on the residuals of turbine 2 or 3, and $d = 2$ for the training on both 2 and 3 combined. Further, the latter is done to investigate the effects of possible over-fitting.

The permutation of training and evaluation time series is shown in Table 2. Each figure shows the accuracy defined by Eq. (19) (ACC) where the time of the failure indicated by the red shaded area, and the precision defined by Eq. (20) (PRC) for selected values of Δ, as indicated by the color bar. Remark that the axis for the ACC is

located on the right hand side of the figures. In order to make the figures of Table 2 more comprehensive, Fig. 6 shows the case of Turbine 1 trained on Turbine 2. The dash line highlights the time under consideration, while the current time is approximative 2 month prior the failure, the accuracy is in the order of 4 days. Better accuracy is achieved as the current time gets closer to the time of failure. On the other hand, from the intersection of the dashed line with the PRC curves, the precision spans from 0.5 to 1 for different values of Δ. For instance, while only 50% of the predictive density function for the failure time is within a 6 h interval around the real failure (PRC = 0.5 for $\Delta = 6$ hours), the precision for intervals higher than $\Delta = 312$ hours is 100%. Like the accuracy, the precision improves as the current time converges to the time of failure.
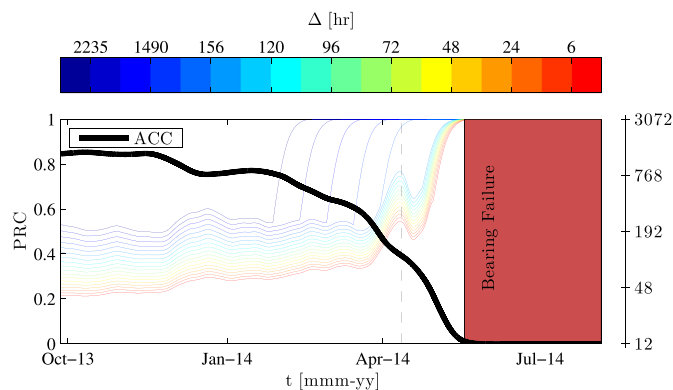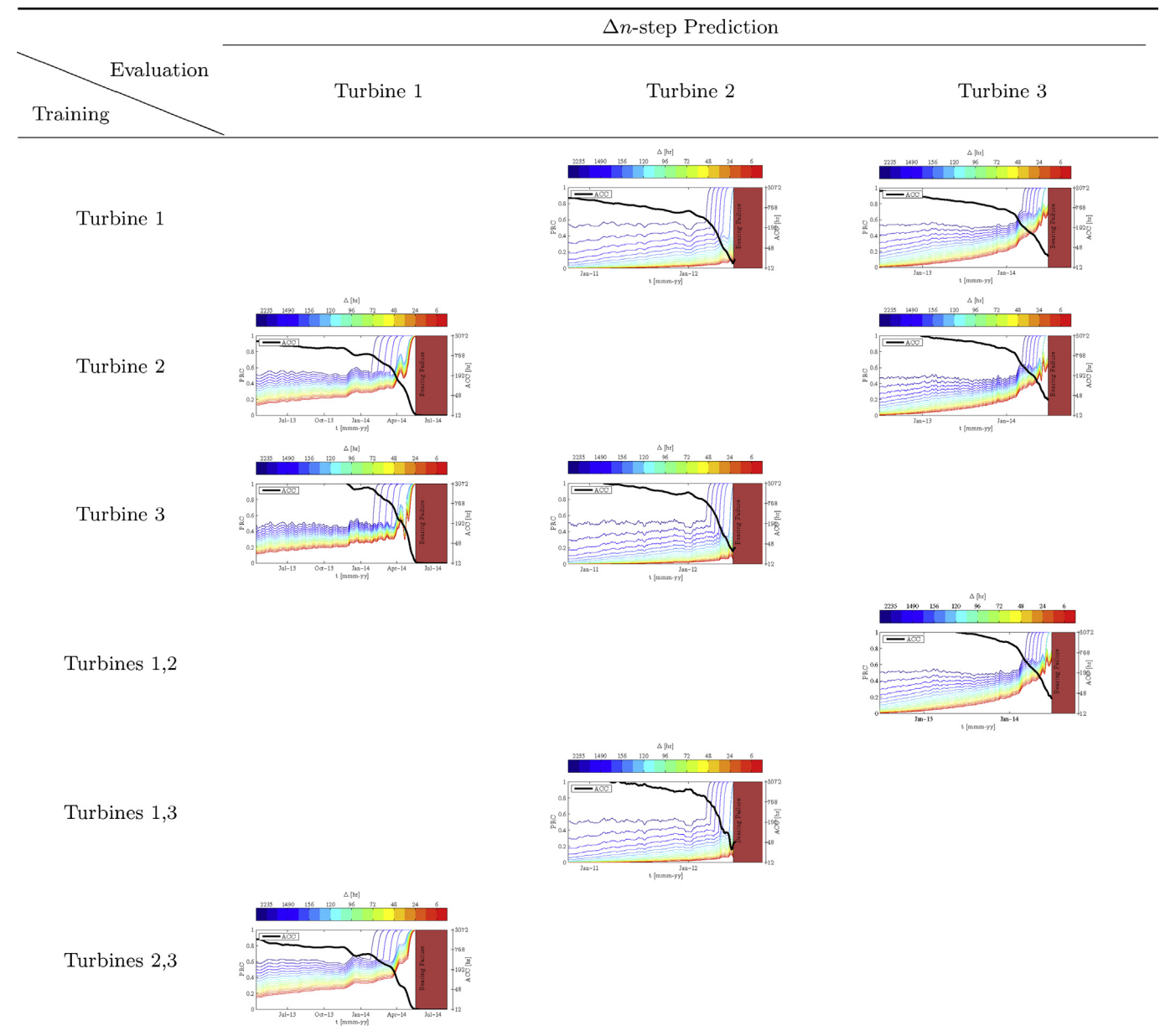
Common for all combinations of turbines and training settings is the trend of the ACC and PRC. The ACC can be separated into two domains: large values far from the monitored failure, and convergences towards low ACC as the failure is picked up by the state prediction. However, notice that only in the case of Turbine 1 does the ACC converge towards zero (full convergence) prior to the fault. Similar domains can be identified for the PRC. For each Δ the time of convergence is different. As Δ specifies the time interval around the failure time, the PRC fully converges when the prediction horizon is close to true failure and in the same order of magnitude as Δ. Again, it is noticed that full convergence is only observed for Turbine 1. Further, as the time to the failure decreases, the PRC associated with Δ become denser, i.e. the predictive distribution $\mathbb{P}(\Delta n | \{E_{k''}\}, s^i, \mathbf{x}_{[1,i]})$ hardens.

Considering the contribution of each turbine during training, it appears that either training on Turbine 2 or Turbine 3 results in a fully convergences of the ACC and PRC for Turbine 1, while it is not observed the other way around. Especially in the case of Turbine 2 it appears that only large values of Δ lead to convergence at all. As the convergence of all turbines resembles the same trend independent of the underlying training data, the conclusion implied is that the rate of convergences is determined by the data under evaluation. From visual inspection, it appears that when training on Turbine 3 more variance in the PRC can be expected. Notice also that the charts of Turbine 3 include more variance overall compared to Turbines 1 and 2. As this is not observed when evaluating Turbine 1 trained on Turbine 2 and vice versa, the increase in variation is attributed to the hyper-parameters updated on Turbine 3.

Training on multiple turbines will increase the robustness of the state prediction but not necessarily increase the precision and accuracy. As the training time series may be composed of trends that are not common for the time series under evaluation, including multiple time series suppresses local variations that will alter the outcome of the prediction. This is most clear when evaluating Turbine 1. As shown in Fig. 1, Turbines 1 and 2 express a dip in the temperature residuals prior to the fault. These dips are observed with different intensity and at different times before the failure. In Table 2 an associated dip is observed in the PRC with different intensity for Turbine 1. However, when combining Turbine 2 and 3 into a training set, the evaluation is more robust against these changes, as the dip in PRC is removed from the chart.

These considerations point towards a strong model dependency on the hyper-parameters, attributed to the data driven nature of the model.

Considering all evaluations subjected to Δ such that PRC ≥ 0.9, an average for the prediction horizon, accuracy, and precision is calculated. The interval around the true occurrence of the fault, as illustrated by the shaded area in Fig. 4, is $\Delta \approx 126$ hours equivalent to ≈ 7 days, with PRC ≥ 0.9, as defined by Eq. (20). The prediction horizon achieved at that point is ≈ 781 hours (≈ 33 days), with ACC ≈ 106 hours (≈ 4 days). In other words, on average, the fault can be predicted 33 days ahead of its occurrence, with 90% of the

Table 2
Accuracy and confidence of selected precisions Δ, for Δn-step predictions.

		Δn-step Prediction		
Evaluation		Turbine 1	Turbine 2	Turbine 3
Training				
Turbine 1				
Turbine 2				
Turbine 3				
Turbines 1,2				
Turbines 1,3				
Turbines 2,3				

Fig. 6. Close-up of the interval around December 1st, 2013, for the PRC and ACC chart in the case of Turbine 1 trained on Turbine 2, as illustrated in Table 2.

predictive distribution $\mathbb{P}(n | \{E_{k''}\}, s^{(i)}, \mathbf{x}_{[1,i]})$ being within the interval $[t_{k''} - \Delta, t_{k''} + \Delta]$, with $\Delta = 7$ days.

6. Conclusion

A state prediction approach has been presented based on the inference of wind turbine bearing temperature residuals and Gaussian processes. Including event data from the individual turbines, it has been shown that prediction of a selected failure event, namely *Bearing Over-temperature*, is possible. Although evaluated on a limited set of time series, the approach has shown promising results, with an averaged time of prediction a month before the actual time of failure with high confidence, and an accuracy and precision in the order of days and a week, respectively.

For the three time series under consideration, it was noticed that only one out of three predictions converges to the true time of

the fault. The others, though close, do not fully converge, with one even showing lower confidence on the partial convergence. Without including more time series, it is not possible to draw certain conclusions. However, it is believed that the discrepancy in the performance can be attributed to the strong data driven nature of the model and the underlying training data. The need to specify the hyper-parameters across all possible states, while training, is a task that requires numerical solutions for the global extrema of highly non-convex cost functions. As such, the calculations become more time consuming when searching for the true optimum, so the model construction becomes a trade-off between computational efficiency and overall model performance.

Exploring the ability of Bayesian inference to abstract and predict wind turbine states conditioned on specific events in more general cases, e.g. using multiple sensors inputs rather than residuals, is a great opportunity for further research.

Acknowledgements

References

[1] J. Nielsen, J. Sørensen, On risk-based operation and maintenance of offshore wind turbine components, Reliab. Eng. Syst. Saf. 96 (1) (2011) 218–229, http://dx.doi.org/10.1016/j.ress.2010.07.007.

[2] K.R. Petersen, E.S. Madsen, A. Bilberg, Offshore wind power at rough sea: the need for new maintenance models, in: 20th EurOMA Conference, 2013.

[3] P.J. Tavner, J. Xiang, F. Spinato, Reliability analysis for wind turbines, Wind Energy 10 (1) (2007) 1–18, http://dx.doi.org/10.1002/we.204.

[4] M. Wilkinson, F. Spinato, M. Knowles, Towards the zero maintenance wind turbine, in: Proceedings of the 41st International Universities Power Engineering Conference, 2006 (UPEC '06), vol. 1, 2006, pp. 74–78, http://dx.doi.org/10.1109/UPEC.2006.367718.

[5] C.A. Walford, Wind Turbine Reliability: Understanding and Minimizing Wind Turbine Operation and Maintenance Costs, Tech. Rep, Sandia National Laboratories, Sandia National Laboratories, Albuquerque, New Mexico 87185 and Livermore, California, 2006, p. 94550.

[6] H. Arabian-Hoseynabadi, H. Oraee, P. Tavner, Failure modes and effects analysis (FMEA) for wind turbines, Int. J. Electr. Power & Energy Syst. 32 (7) (2010) 817–824, http://dx.doi.org/10.1016/j.ijepes.2010.01.019. URL, http://www.sciencedirect.com/science/article/pii/S0142061510000281.

[7] J.B. Gayo (Technical Unit Manager), Reliability-focused Research on Optimizing Wind Energy System Design, Operation and Maintenance: Tools, Proof of Concepts, Guidelines & Methodologies for a New Generation, Project Final Report, ReliaWind, 2011.

[8] S. Arabgol, H.S. Ko, S. Esmaeili, Artificial neural network and ewma-based fault prediction in wind turbines, in: Proceedings of the 2014 Industrial and Systems Engineering Research Conference, 2014.

[9] A. Kusiak, A. Verma, A data-driven approach for monitoring blade pitch faults in wind turbines, Sustain. Energy, IEEE Trans. 2 (1) (2011) 87–96, http://dx.doi.org/10.1109/TSTE.2010.2066585.

[10] A. Kusiak, W. Li, The prediction and diagnosis of wind turbine faults, Renew. Energy 36 (1) (2011) 16–23, http://dx.doi.org/10.1016/j.renene.2010.05.014. URL, http://www.sciencedirect.com/science/article/pii/S0960148110002338.

[11] O. Bennouna, N. Heraud, Z. Leonowicz, Condition monitoring amp; fault diagnosis system for offshore wind turbines, in: Environment and Electrical

Engineering (EEEIC), 2012 11th International Conference on, 2012, pp. 13–17, http://dx.doi.org/10.1109/EEEIC.2012.6221389.

[12] K. Fischer, F. Besnard, L. Bertling, Reliability-centered maintenance for wind turbines based on statistical analysis and practical experience, Energy Convers. IEEE Trans. 27 (1) (2012) 184–195, http://dx.doi.org/10.1109/TEC.2011.2176129.

[13] J. Herp, E.S. Nadimi, Wind turbine fault detection based on artificial neural network analysis of scada data, in: EWEA Offshore 2015 Conference Proceedings, 2015.

[14] J. Herp, N.L. Pedersen, E.S. Nadimi, Wind turbine performance analysis based on multivariate higher order moments and bayesian classifiers, Control Eng. Pract. 49 (2016) 204–211, http://dx.doi.org/10.1016/j.conengprac.2015.12.018. URL, http://www.sciencedirect.com/science/article/pii/S096706611530068X.

[15] P. Odgaard, J. Stoustrup, M. Kinnaert, Fault-tolerant control of wind turbines: a benchmark model, Control Syst. Technol. IEEE Trans. 21 (4) (2013) 1168–1182, http://dx.doi.org/10.1109/TCST.2013.2259235.

[16] Y. Yan, J. Li, D.W. Gao, Condition parameter modeling for anomaly detection in wind turbines, Energies 7 (5) (2014) 3104–3120, http://dx.doi.org/10.3390/en7053104. URL, http://www.mdpi.com/1996-1073/7/5/3104.

[17] A. Kusiak, A. Verma, Analyzing bearing faults in wind turbines: a data-mining approach, Renew. Energy 48 (0) (2012) 110–116, http://dx.doi.org/10.1016/j.renene.2012.04.020. URL, http://www.sciencedirect.com/science/article/pii/S0960148112002613.

[18] M. Bach-Andersen, B. Rømer-Odgaard, O. Winther, Flexible non-linear predictive models for large-scale wind turbine diagnostics, Wind Energy http://dx.doi.org/10.1002/we.2057. URL http://dx.doi.org/10.1002/we.2057.

[19] A. Kusiak, Z. Zhang, A. Verma, Prediction, operations, and condition monitoring in wind energy, Energy 60 (0) (2013) 1–12, http://dx.doi.org/10.1016/j.energy.2013.07.051. URL, http://www.sciencedirect.com/science/article/pii/S0360544213006579.

[20] F.P.G. Márquez, A.M. Tobias, J.M.P. Pérez, M. Papaelias, Condition monitoring of wind turbines: techniques and methods, Renew. Energy 46 (0) (2012) 169–178, http://dx.doi.org/10.1016/j.renene.2012.03.003. URL http://www.sciencedirect.com/science/article/pii/S0960148112001899.

[21] Y. Vidal, C. Tutivén, J. Rodellar, L. Acho, Fault diagnosis and fault-tolerant control of wind turbines via a discrete time controller with a disturbance compensator, Energies 8 (5) (2015) 4300–4316, http://dx.doi.org/10.3390/en8054300. URL, http://www.mdpi.com/1996-1073/8/5/4300.

[22] Z. Gao, X. Liu, M.Z.Q. Chen, Unknown input observer-based robust fault estimation for systems corrupted by partially decoupled disturbances, IEEE Trans. Ind. Electron. 63 (4) (2016) 2537–2547, http://dx.doi.org/10.1109/TIE.2015.2497201.

[23] X. Liu, Z. Gao, Robust finite-time fault estimation for stochastic nonlinear systems with brownian motions, J. Frankl. Inst. (2016), http://dx.doi.org/10.1016/j.jfranklin.2016.08.018. URL, http://www.sciencedirect.com/science/article/pii/S0016003216302939.

[24] J. Herp, M.H. Ramezani, E.S. Nadimi, Dependency in State Transitions of Wind Turbines - Inference on Model Residuals for State Abstractions, (accepted in IEEE Transactions on Industrial Electronics).

[25] S. Dey, P. Pisu, B. Ayalew, A comparative study of three fault diagnosis schemes for wind turbines, Control Syst. Technol. IEEE Trans. 23 (5) (2015) 1853–1868, http://dx.doi.org/10.1109/TCST.2015.2389713.

[26] P. Odgaard, J. Stoustrup, A benchmark evaluation of fault tolerant wind turbine control concepts, Control Syst. Technol. IEEE Trans. 23 (3) (2015) 1221–1228, http://dx.doi.org/10.1109/TCST.2014.2361291.

[27] J. Herp, M.H. Ramezani, State transition of wind turbines based on empirical inference on model residuals, in: Proceedings of the 2015 Conference on Research in Adaptive and Convergent Systems, RACS 2015, Prague, Czech Republic, October 9-12, 2015, 2015, pp. 32–37, http://dx.doi.org/10.1145/2811411.2811509.

[28] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.

[29] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman and Hall,CRC, 2014.

[30] R. von Mises, Mathematical Theory of Probability and Statistics, 1964.

[31] M. Buhmann, Radial Basis Functions: Theory and Implementations, Cambridge Monographs on Applied and Computational Mathedmatics, Cambridge University Press, 2003.