Review for:

Regression-based Main Bearing Load Estimation
MS No.: wes-2022-75
MS type: Research article

Reviewer: Edward Hart

General comments

This paper considers the problem of estimating wind turbine main bearing loads based on deflection and/or strain measurements taken within the drivetrain. While a valuable endeavour, the paper is quite significantly flawed in its current form. Extensive revisions and additions would therefore be required before I could recommend it for publication. In this regard, I will provide some general feedback, followed by a detailed breakdown of my comments on specific parts of the manuscript.

1) Language and quality of writing – the quality of writing is often poor and needs improving, the manuscript therefore requires extensive revisions throughout.

2) Clarity – The paper lacks a clear and consistent narrative and structure. As a result, the reader is often left unsure as to what each section is trying to achieve, and exactly what the aims are. For example, in Section 3, a number of methods/methodologies/equations are presented one after the other without proper set up. It therefore quickly become confusing trying to glean whether each is a separate method to be tested, or whether two or more are to be used together. A clear and consistent narrative and proper signposting would help here, as well as the possible addition of a flow diagram to provide an overview of the study.

3) Rigour of statistical analyses – As outlined in one of my detailed comments below, I am uncomfortable with the approach used for training and testing within this study. Furthermore, there is no discussion of a suitable cross-validation framework for model fitting. The applied non-parametric methods are treated very much like black-boxes, which I also see as potentially problematic. Confidence intervals were mentioned in relation to the GP approach, but not actually applied or investigated (similarly for other methods). At the very least I'd like to see a more careful discussion of these techniques in an updated manuscript. While I'd be glad to see an improved approach to training and testing, I will not insist on it if the description and discussion around the statistical models and their results can be improved.

4) Value/applicability of the outcomes – There is certainly a lot of value in methods with which main bearing loads could be estimated accurately in operational wind turbines based on inexpensive sensors. However, with the paper in its current form, it is not at all clear if/how this work may be contributing to this eventual goal. More detailed discussion therefore needs to be added, which directly tackles the implications of this work for the end goal. For example, the good results seen here require the models to be trained on full size test-bench data. Would the same training data be expected to allow for accurate prediction in an operational wind turbine? If not, then how would a training set be developed and validated? In the latter case, it would also be important to understand how much training data is required to bring model predictions within an acceptable level of accuracy/confidence. Many of these questions will not have answers at this stage, but discussion along these lines would allow the current work to help direct and drive future efforts through which they may be concretely tackled.

I will emphasise that I am not necessarily insisting that every single point made above be bottomed-out now. But a revised manuscript should certainly seek to either address or, at least, discuss these points in some form.

Specific comments

L5 “The premature failure of wind turbines due to unknown loads leads to a reduction in competitiveness compared to other energy sources.” Source?

L7 “Load monitoring systems can make a significant contribution to understand and prevent such failures.” Source?

L17 “overproportional higher non-torque loads” poor English, similar issues exist throughout the manuscript

L23 “Nevertheless, all these damage mechanisms are load-driven” are they? What about contamination, electrical current, poor maintenance…

L24 “A direct measurement of main bearing loads is not possible as there are no force transducers available for this load range.” Is that true? What about recent NREL work with SKF, shaft moment measurements etc. Please see recent NREL technical reports and papers on these topics.

L28 “dynamic” has a specific meaning, do you mean non-steady?

L32 “accurately documented which operational loads have led to main bearing failures” I don’t believe that information is currently known or necessarily available, hence this statement seems misleading

L35 “mainly logs its performance parameters” I’m not sure if I’d refer to these as performance parameters, maybe operational or control parameters?

L37 “Thereby mean, maximum and minimum values are recorded covering a time span of up to ten minutes.” Along with standard deviations.

L39 “Pagitsch et al (Pagitsch et al., 2020)” Please use cite, rather than citep, in latex to just reference in-line rather than in parenthesis.

L43 “But this gives only partial information on the complete main bearing loading what complicates a later detailed understanding of harmful load situations.” Poor English, please revise.

L46 “Various sensor technologies are available to measure main bearing related loads with additional sensors” Poor English again. I believe that the entire manuscript needs to be carefully proofread and improved regarding the language used.

L46 “Hereby, the measurement of strains has been a method widely used especially in prototype validation and certification of wind turbines. Regarding main bearing loads, strain gauges are typically applied to the main shaft after the first main bearing.” Sources? You also haven’t described drivetrain layouts, so the reader may not be aware that turbines may have more than one main shaft bearing.

L61 “has not yet been taken place” language again

L63 you don’t reference any NREL/SKF work at this stage?

L70 “while the use of linear regression models does have the advantage of a more robust behaviour due to main bearing clearance uncertainties (Loriemi et al., 2022).” Do you mean “in the presence of main bearing clearance uncertainties.”

Fig 2. Please provide some labelling for the different force components to allow easier comparison with Equations 1-5.

Equations 1-5. Why do the $F_{z,GS}$ terms not include a cos(alpha) term?

L109 “In the preparation of the tests, the TS and GS were measured on component test benches and three-dimensional stiffness curves were recorded. This allows to calculate the reaction forces of these components.” What was the accuracy here? Is this a possible source of error/uncertainty?

L115 “Before testing all full bridges have been shunt calibrated.” What is the accuracy here? Is this a source of possible error?

L120 “Additionally to the strain gauges, sixteen eddy current sensors” no proper background was provided for eddy current sensors.

Equation 6. Has the symbol alpha been used for two different purposes? Here and in Fig 2?

L132 “Nevertheless, in this study all available sensors are used to derive a more averaged measurement.” This sentence is unclear. Is “more averaged” a good thing? Also, how does Equation 6 clarify that one senor could be removed while displacements could still be determined? Please explain clearly (I am guessing it’s due to there being 5 unknowns, and the fact that these unknowns are linear combinations of the measured quantities).

L134 “dynamic” -> “non-steady”.

L136 Please provide reference for IEC 61400, and include the specific part, i.e. 61400-1

3.2 Test Plan - This training/testing plan appears to be fundamentally flawed. From a data science perspective, I can think of no good reason to split testing and training data along these lines. Indeed, the conscious construction of training sets which deliberately leave out a certain set of conditions is poor practice which, for some applications, will lead to over-fitting to the conditions seen during training. This could in-turn lead to poor results during testing/application phases. Note, even if good performance is observed from this split of training and testing, it remains the case that this is poor practise and not to be recommended. When developing a framework within which to fit and test statistical regression models, the split between testing and training data is certainly very important, and should ideally be done in a way which reflects the context in which the developed models would then be applied. I believe a superior approach here would be to generate a training set which is composed of data from all wind classes, which is then tested on a separate (disjoint) dataset which is also generated such that it covers all wind classes. There would therefore be a complete split between training and test data, but without consciously leaving out certain conditions from the training data. The development of such models would also ordinarily involve some form of cross validation (e.g. k-fold). This appears to not be considered here.

L160 “As described before, this physics-based approach is subject to uncertainties” isn’t it ‘error’ rather than uncertainty? Or maybe it is best described as an ‘approximation’.

Equations 6-9 and 10-13: If the sensors need calibrating then the measured ‘strains’ are known to be imperfect. Equations 6-9 are therefore not actually useful as far as I can tell, since any result you see here may be wildly different for different sensors installed on a different turbine. Effectively, only after calibration would one hope that those measurements can tell us anything useful. If this is the case then I'm not sure if plotting results for Eqs 6-9 is meaningful.

L180 "Using test bench measurements as training data different six different regression models are trained for main bearing load estimation." I find this a little confusing. You are predicting main bearing loads from displacements, but the main bearing loads are themselves estimated from a force balance which includes terms which are estimated via an LVDT.

L204 The non-parametric regression models should be described at least a little when introduced. It is important that these techniques are not just treated as black boxes. You state "Another advantage of this algorithm is the comparatively good robustness against noise, which may be required due to a partially bad signal-to-noise-ratio of displacement signal". "Partially bad'' doesn't make sense, plus you don't mention anything about these signal characteristics anywhere else in the paper. I would think that the characteristics of these signals, and how that can vary under different conditions or ambient temperatures etc, would be critical to the overall discussion.

L209 "In addition to the achievable good accuracy of this algorithm, there is the possibility to provide confidence intervals for the estimated loads." All statistical models (including linear regression) can have confidence intervals produced for them. The GP approach simply produces them more readily, as they are essentially 'built in'. Also GPs and FNNs are highly non trivial techniques which are suitable for some problems and not for others. Blind testing without careful cross-validation and appropriate training may well produce poor results, even if better outcomes are possible. I would prefer a more careful treatment of these different methods, at least having them introduced a little more fully.

L240 "This error is massively reduced by the calibration procedure of applying single loads to the DUT. The approach of using the training data (Table 1) to derive best fitting coefficients (cf. equation 10-13) results in the best performing regression model, which uses only strain signals and outperforms almost every displacement-based regression model." This sentence appears to describe the Calibrated and Regression models at the same time, please review as I can't make sense of it. Related to that, what is the "FITTED" model? If that's the one you're referring to, it doesn't do better everywhere relative to the other models (e.g. F_{yFB})

L254 "One finding was the strong deviation of the analytic and calibrated strain-based regression model. Comparing the analytically calculated coefficients with the regression coefficients of the calibrated model, it can be stated that deviations mainly derive from the young modulus." I'm not sure if this is surprising or that valuable an observation. These sensors are relative, rather than absolute, so we expect to have to calibrate right?

Table 2 – please discuss the magnitude of RMSE errors in an absolute sense. You only really discuss them relative to the different models. If I understand correctly, the RMSE is shown as a % (relative to the load-range - what exactly is that?). But that indicates that the RMSE errors are >100%, or sometimes close to 200% of the load range. Isn't that pretty bad?

L266 "The use of displacement signals for load estimation has been shown to provide an equivalent accuracy as using strain signals." There's caveats to this aren't there, i.e. you need to train a model and so have access to appropriate data on which to do that.

L275 "A technically feasible variant in this application would be the use of axial displacement signals on the rotor flange, which have good signal-to-noise-ratio due to the relatively large shaft deflection” Again, you mention signal characteristics in discussions, but I feel this aspect of the study could do with its own subsection and some figures etc.