Response to Referee 2

We greatly appreciate the time taken by the referee to read our manuscript. We have taken into consideration and addressed all comments, questions, and suggestions from the reviewer, and we feel that the revised manuscript is now substantially stronger as a result. Changes made to the text at the request of the reviewer have been highlighted in red in the revised manuscript. In the following, reviewer comments are repeated in italics and our responses are provided in the regular sections of text.

Point Comments

1. *Page 2, Line 26: "Damiani et al. (2018) performed a detailed analysis of a single wind turbine, noting that negative yaw offsets tended to increase fatigue loading more than positive yaw offsets (although it should be noted that these results were specific to the turbulence seeds used in the study)." The added parenthetical remark does not provide the necessary caveat. Damiani et al. (2018) specifically cautioned that the influence of the incident conditions "make generalization more difficult." The authors should state that Damiani et al. (2018) cautioned against the general statement that "negative yaw offsets tended to increase fatigue loading more than positive yaw offsets."*

We agree that this was too general of a statement. We have revised the manuscript on P2L26-31 to clarify that Damiani et al. (2018) stressed against generalizing the findings of their single turbine study. We have also now added references to studies of wake steering effects on downstream turbines. In particular, we point to Figures 2 and 4 in (López et al., 2020) and Figure 3 in (Zalkind and Pao, 2016).

2. *Page 2, Line 37: I am still concerned with the phrasing "remarkably accurate in power prediction." This seems to highlight that wake modeling for power predictions is a concluded matter. I don't believe this is the case, as wake models can exhibit high predictive error for many utility-scale wind farm applications, e.g. time-varying conditions, the presence of complex terrain, strong stratification, etc., and this statement is counter to the community calls for research [e.g. 1, 2].*

We agree that this phrasing was somewhat overzealous. We removed this sentence to avoid causing confusion. We now only claim that engineering wake models have dubious accuracy when predicting fatigue loading on P2L38-40.

3. *Page 3, Line 73: "We use the scikit-learn Gaussian Process implementation (Pedregosa et al., 2011), which is a well-validated open-source project." It's useful that the authors have added this, but specifically, I would like to see: 1) a comparison of the GP fit to the training data; 2) a GP prediction (out of sample of the training data) compared to the simulated power and loads for an out of sample set of the design variables.*

These are important points and we have consequently conducted a leave-one-out analysis to assess the accuracy of the single-fidelity (SF) and multifidelity (MF) GP models, excluding one point at a time to compare the prediction of the GP model to the observed value. Excluding some points makes their associated predictions outside of the training range. The results and a corresponding discussion were added to a new appendix at the end of the revised paper, also shown in Figures 1 and 2 at the end of this document. Generally, there is slightly more error incurred by the prediction of the MF GP than the SF GP. We reference the results in the appendix on P14L343-P15L344.

4. *Page 5, Line 118: Many readers of Wind Energy Science will not be familiar with Pareto dominance, and this article should be self-contained. Please add one or two sentences defining Pareto dominance before referring to the paper or remove its use from the manuscript.*

We have added a brief description of what Pareto dominance means on P5L119-120.

5. *Page 9, Line 221: "These time parameters were justified by comparing power and loading computed over time intervals of 600-900 s and 900-1,200 s, resulting in relative differences of only 2.6% for power and 4.2% for loading when $\gamma = (15; 0)$." These differences seem nontrivial. Have the authors tested how sensitive the Pareto set is to these finite-time averages? Does the optimal yaw change 0.1 degrees or 10 degrees? Especially given how sensitive the flow physics interpretation is to the particular wind conditions, I am interested in hearing more about this.*

We agree that the temporal convergence of the underlying simulation is an important concern. To further test this convergence, we performed longer simulations at each of the sample points during the optimization, and Figure 3 at the end of this document compares the Pareto fronts found using time intervals of 600-1,200 seconds and 1,200 seconds-1,800 seconds. Although the shape of the Pareto front does change slightly, the hypervolume of the Pareto fronts discovered by the single-fidelity and multifidelity optimizations each differed by less than 0.5% between the two analysis periods. We have added text summarizing this further analysis on P9L223, and we have also added to the conclusions on P19L410-412 that care should be taken when applying this method to ensure convergence of the Pareto set with respect to convergence of the underlying simulation.

6. *Eq. (32): Now that you have clarified that the choice of the objective function (especially the -10 in the loads function) was ad hoc to have a negative value, I am wondering about the effect of your objective function here. Is the particular quantity of the objective function entirely arbitrary in your framework or will it impact your Pareto front and/or your exploration/exploitation tradeoff? Perhaps the mean of the objective does not matter but it is the sensitivity with respect to the design variables which matters? Simply stated, a reader will want to know: how do I pick an objective function if I want to apply this proposed method? Should it depend on the turbine model, farm geometry, etc.?*

We have clarified on P11L269-271 that, because the EHVI is an area produced by the two objectives, we do not expect different values in this scaling function to affect the results of maximizing the acquisition function, provided that all sampled objective values are always less than the associated reference value.

We note that this scaling must be done in *ad hoc* manner, as there is no "optimal" scaling of the two functions that we are aware of, and having the objectives all have the same goal of minimization simplifies the problem formulation. The scaling should be selected so that the functions are of similar magnitude and will never change sign. Since the expected hypervolume improvement acquisition function is based on the area formed by the two objectives, we do not expect the magnitude of the individual objectives to have a great influence on the optimal choice. For example, we internally confirmed that using a different load scaling (i.e., $L = \hat{L}/10^6 - 15$) results in the same solution of which set of yaw offsets to sample next in a multifidelity case.

7. *It is reasonable to ask whether the presented results actually show that the multifidelity approach is better than the single fidelity. For minimizing EHVI, it seems that multifidelity is better (al-*

though the noise is larger for multifidelity and if you extrapolate the trend lines it would seem to suggest the blue line will drop below the orange with more evaluations). For the end objectives of power and loads, the multifidelity is only faster to estimate minimum loads, it is actually slower for estimating maximum power. One also has to recognize that these are the results where the authors have intentionally created an artificial loads model such that the multifidelity is better than the single fidelity (Page 11, Line 286). Does this bias the results? I would appreciate for the authors to confront this more directly. The paper is clearly a novel, significant contribution already based on the multiobjective optimization approach. But can it be concluded that multifidelity is superior to single fidelity? Perhaps this is a call to action for better loads modeling more than anything.

We agree that our original draft reads as if the power converged faster in the single-fidelity case, which would be an unintuitive result. We removed text on P12L310-313 that had erroneously claimed the single-fidelity optimization optimal power converged faster than the multifidelity counterpart, when, in fact, the single fidelity approach had only converged to the *approximate* optimal solution faster. Close inspection of Figure 3 in the paper shows that the power for the single-fidelity approach converges to the best solution later than in the multifidelity case.

We also stress that the most important convergence metric – indeed, the ultimate goal of the present approach – is to achieve convergence of the hypervolume spanning multiple objectives. That is, we can only claim true convergence once *all* objectives (i.e., both loading and power) have been optimized. In this sense, the single-fidelity approach is less efficient because it takes longer than the multi-fidelity approach to converge for all objectives.

Nevertheless, as with the Damiani *et al.* (2018) study, we caution against generalizing the present results to all wind farm optimization sutides. We have added text to the conclusions on P18L386-387 confronting the possibility of bias and calling for this approach to be applied to more wind farm layouts to assess if the single-fidelity approach consistently finds the optimal power production faster than the multifidelity approach.

Appendix

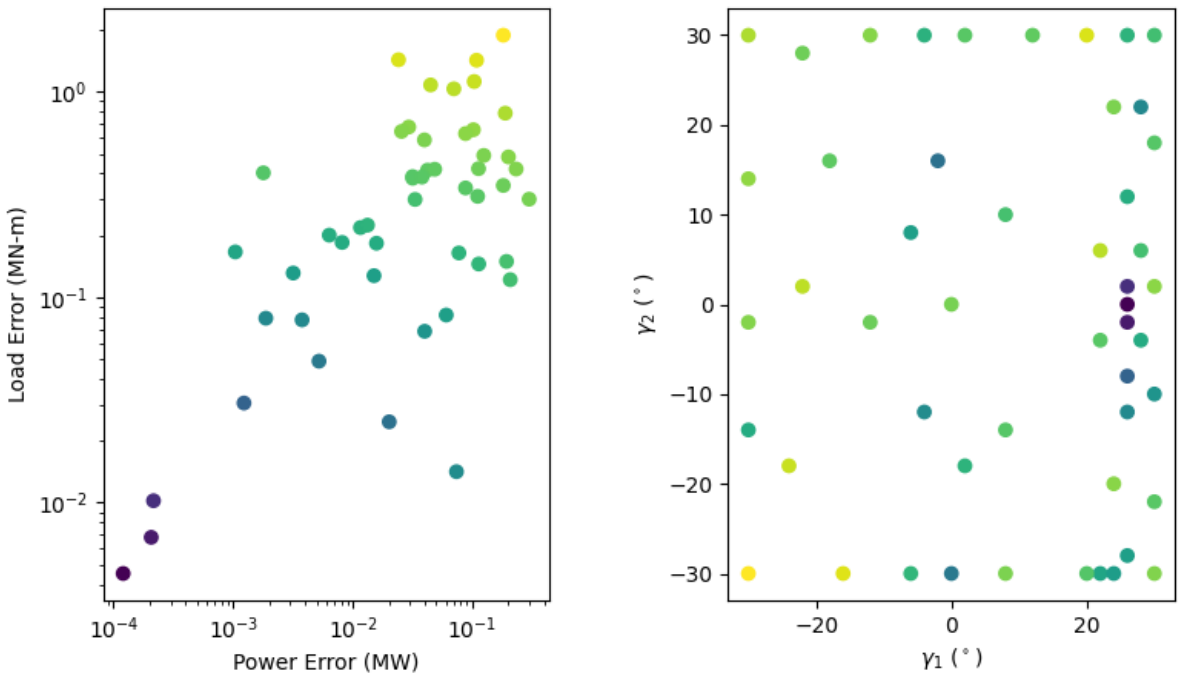

Figure 1: Results of single-fidelity LOO. The left panel shows the leave-one-out prediction errors associated with power and loading, and the points are colored by the sum of both errors. The same points are plotted in the right panel, showing their associated γ_1 and γ_2 values.

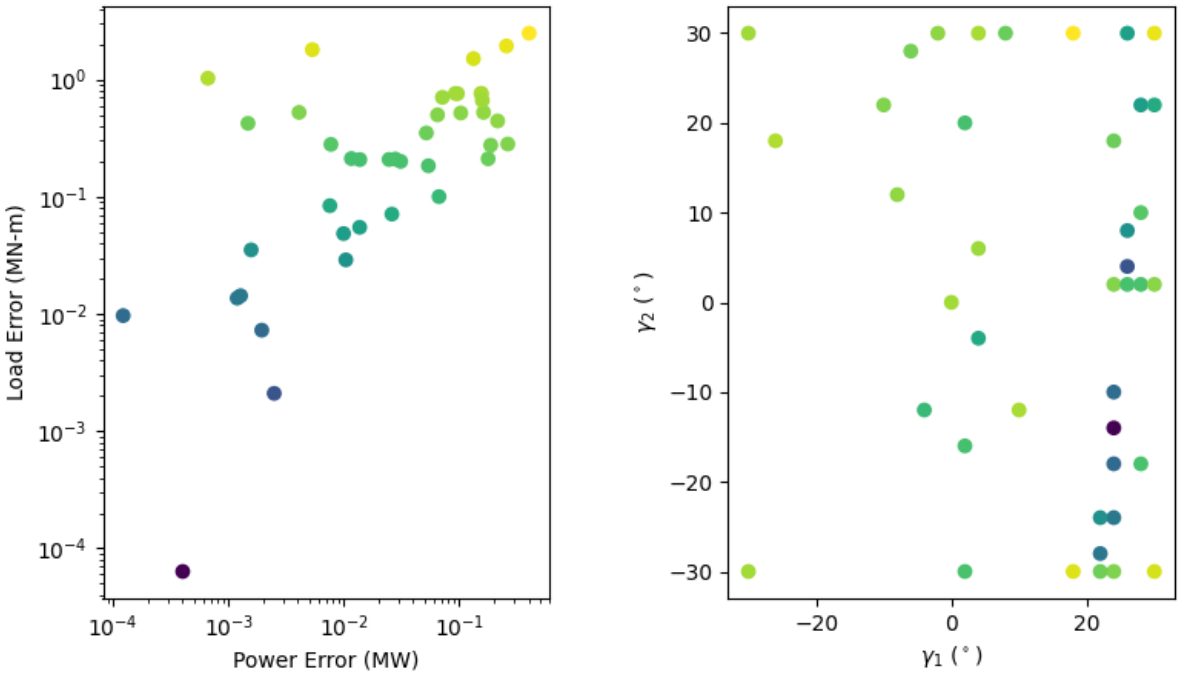

Figure 2: Results of multifidelity LOO. The left panel shows the leave-one-out prediction errors associated with power and loading, and the points are colored by the sum of both errors. The same points are plotted in the right panel, showing their associated γ_1 and γ_2 values.

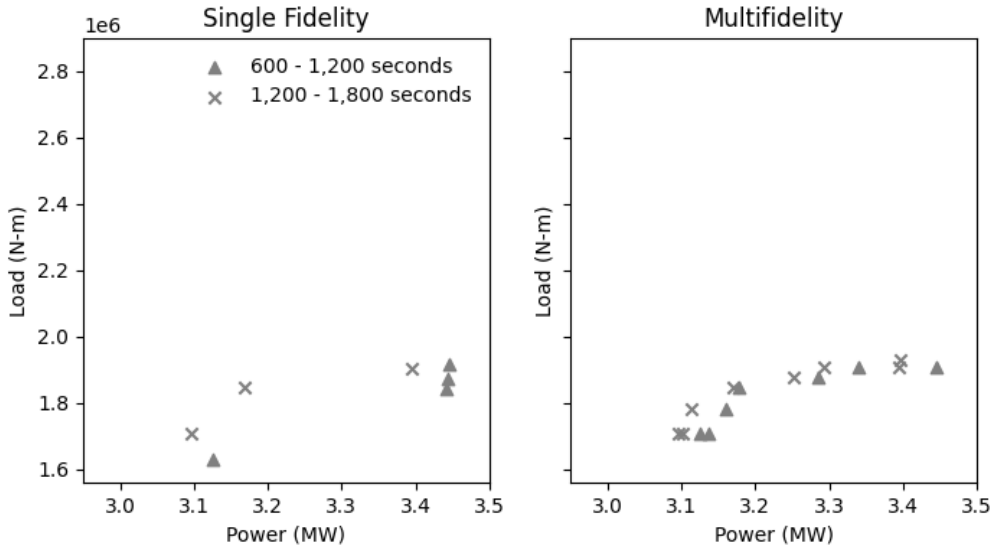

Figure 3: Pareto fronts discovered by the single-fidelity and multifidelity optimization algorithms, computed using the original time wind (600-1,200 seconds, triangles) and a new time window (1,200-1,800 seconds, crosses). The plot in the left column results associated with the single fidelity optimization and the plot in the right column shows the results associated with the multifidelity optimization.