Responses to Reviewers' Comments for Manuscript 10.5194/wes-2026-54

# Convolutional versus graph-based surrogate models for inter-farm wake prediction using multi-fidelity transfer learning

Addressed Comments for Publication to

WES Wind Energy Science

by

Jens Peter Schøler, Frederik Peder Weilmann Rasmussen, M. Paul van der Laan, Alfredo Peña, and Pierre-Elouan Réthoré

## Document Overview

This document contains the authors' responses to the reviewers' comments for manuscript wes-2026-54, "Convolutional versus graph-based surrogate models for inter-farm wake prediction using multi-fidelity transfer learning." The document is structured as follows:

- **Section 1:** Responses to Reviewer 1's comments
- **Section 2:** Responses to Reviewer 2's comments

# Authors' Response to Reviewer 1

**General Comments.** This manuscript presents a rigorous and timely comparison of ARU-Net and GNO surrogate models for inter-farm wake prediction using multi-fidelity transfer learning. The study addresses a critical challenge in offshore wind energy as farm clusters become denser. The key findings offer valuable and actionable insights for surrogate model selection in wind energy applications. The paper is well-structured and the experiments are reproducible. Addressing the requested clarifications on computational efficiency, out-of-distribution robustness, and certain methodological details will further strengthen this solid contribution. I recommend acceptance pending major revisions.

**Response:**

We thank the reviewer for the positive assessment and constructive comments. In response, we have revised the manuscript to improve the Introduction, clarify methodological choices, soften the interpretation of transfer-learning gains, and expand the discussion of computational timing, grid sensitivity, and model limitations.

## 1. Introduction

**Comment 1**

The transition from the motivation for multi-fidelity transfer learning to the side-by-side introduction of CNN and GNN architectures (Page 3, Line 80) is somewhat disjointed. It would be helpful to state that CNNs and GNNs represent the two principal architectural paradigms currently being explored for implementing such multi-fidelity surrogates.

**Response:**

We have added a bridging sentence before introducing the CNN and GNN architectures to clarify that these represent two principal architectural paradigms for implementing multi-fidelity surrogate models in this context.

**Comment 2**

The topic about inter-farm wake interaction for the wind farm clusters should be reviewed deeply, such as the literature reported in Journal of Cleaner Production 2023, 396: 136529 and Energy Conversion and Management 2022, 267, 115897.

**Response:**

We have added a short discussion of the two suggested studies to the Introduction, highlighting their findings on atmospheric stability, terrain effects, offshore cluster wakes, and their relevance to fast inter-farm wake modelling.

**Comment 3**

Please ensure that all abbreviations are expanded upon first use. For instance, "SciML" in Section 2.1 should be written as "Scientific Machine Learning (SciML)" on its first occurrence.

**Response:**

We have checked the manuscript for abbreviations and ensured that discipline-specific abbreviations are expanded upon first use. In particular, "RMSE" has been expanded to "root mean squared error (RMSE)" in the Abstract, "SciML" has been changed to "scientific machine learning (SciML)" in Section 2.1, and "ReLU" has been expanded to "rectified linear unit (ReLU)" at its first occurrence.

## 2. Methodology

**Comment 4**

The manuscript states that PLayGen generates four distinct layout types for the low-fidelity dataset. However, it appears that the high-fidelity RANS-AWF dataset is restricted predominantly or exclusively to cluster layouts. Could the authors clarify the rationale for this restriction? Is it due to computational constraints, or is the cluster layout considered sufficiently representative for the physics corrections targeted by the high-fidelity data? This clarification is important for understanding the generalizability of the fine-tuned models.

**Response:**

The RANS-AWF dataset is repurposed from Rasmussen et al. (2026), where high-fidelity simulations were generated for cluster-like inter-farm wake configurations using random farm polygons and Poisson-disk turbine placement. We have clarified this in the dataset description and added that the fine-tuned models should therefore be interpreted as most directly representative for this layout class.

**Comment 5**

The ARU-Net is implemented in PyTorch, whereas the GNO is built on JAX/Jraph. Could the authors comment on whether this difference in underlying frameworks could introduce any biases in the reported wall-clock training times (Table 4)? For instance, data loading overhead or graph construction routines may differ in efficiency, potentially affecting the comparison of LoRA versus full fine-tuning efficiency.

**Response:**

We have clarified in the discussion of Table 4 that the wall-clock times are not intended as a framework-normalised benchmark between ARU-Net and GNO, since the models use different software stacks and data pipelines. The ARU-Net is implemented in PyTorch, while the GNO uses JAX/Jraph/Flax together with PyTorch Geometric for graph construction and data loading. We therefore interpret the timing results primarily within each architecture, comparing LoRA, freezing, full fine-tuning, and training from scratch using the same implementation framework.

**Comment 6**

The authors clamp velocities exceeding $U_\infty$ to $U_\infty$ prior to the log-deficit transformation for the ARU-Net. This effectively forces the ARU-Net to ignore regions of flow acceleration caused by blockage effects. Could this design choice partially explain the observed differences in F1 scores and RMSE between the two models? A brief discussion of whether the GNO's ability (or inability) to predict such speed-ups affects the practical utility of the respective models would be insightful.

**Response:**

We have added a discussion noting that the ARU-Net target transformation clamps velocities above $U_\infty$ before applying the logarithmic wake-deficit transform, so local speed-ups are not preserved in the ARU-Net target representation. We also clarify that this is expected to have limited influence on the threshold-based wake metrics, since the wake mask is defined using $\delta_w \geq 10^{-3}$, whereas clamped speed-up and unwaked regions correspond to the much smaller floor value $\varepsilon_0 = 10^{-5}$. The main relevance of this design choice is therefore for full-field error comparisons and for applications where blockage-induced acceleration is of interest.

## 3. Results and Discussion

**Comment 7**

The manuscript states that the GNO benefits "only marginally" from transfer learning. While the relative improvement is indeed smaller than that observed for the ARU-Net, Table 4 shows a reduction in RMSE from 0.00696 (trained from scratch) to 0.00558 (LoRA fine-tuned), which represents a ∼20% relative improvement. In the context of wind farm energy yield assessment, this may not be negligible. The authors should consider softening the language to reflect that while the relative gain is modest compared to the ARU-Net, pre-training still yields the best absolute performance for the GNO.

**Response:**

We agree that describing the GNO improvement as "only marginal" was too strong, since the best LoRA fine-tuned GNO reduces the validation RMSE from 0.00696 to 0.00558 $\mathrm{m\,s^{-1}}$ compared with training from scratch. We have softened the wording in the Abstract and Conclusion, and added a clarification in the discussion of Table 4, noting that the GNO gains are smaller than for the ARU-Net but still measurable, with the best absolute GNO performance obtained with transfer learning.

**Comment 8**

The observation that LoRA fine-tuning offered no significant wall-clock time savings despite the substantial reduction in trainable parameters is counterintuitive. Could the authors elaborate on the likely bottleneck? Was the training time dominated by data loading, graph construction, or the forward pass through frozen layers?

**Response:**

We have expanded the discussion of Table 4 to clarify that the wall-clock time is not dominated solely by the number of trainable parameters. Even when LoRA is used, the frozen layers must still be evaluated during the forward pass, and for the GNO, additional overhead can arise from graph batching, graph construction, and data loading. Thus, LoRA primarily reduces the cost of gradient updates and optimiser states, while much of the per-epoch computation remains unchanged. This likely explains why the large reduction in trainable parameters does not lead to a proportional reduction in wall-clock time.

**Comment 9**

The ARU-Net performance degrades noticeably when interpolated from the CNN grid to the GNN grid. Please specify the interpolation method used (e.g., bilinear, bicubic). Additionally, comment on whether this sensitivity to grid resolution poses a practical concern for deployment scenarios where query points may not align with the training raster.

**Response:**

We have clarified in the caption of Table 5 that bilinear interpolation is used when mapping ARU-Net predictions from the native CNN grid to the GNN grid. We have also added a short discussion noting that the observed degradation on the GNN grid reflects both interpolation error and the fixed-raster nature of the ARU-Net. This sensitivity is a practical limitation when predictions are required at arbitrary query locations that do not align with the training grid, whereas the GNO can evaluate arbitrary probe locations directly.

## 4. Conclusion

**Comment 10**

Further research directions are suggested.

**Response:**

We have expanded the Future work subsection with an additional research direction on graph-transformer models. Specifically, we note that graph-transformer models may provide a flexible attention-based mechanism for capturing long-range inter-turbine interactions while retaining the layout adaptability of graph-based surrogates.

## Recommendation

**Comment 11**

Ensure all acronyms (e.g., SciML, PEFT, FiLM, SE, AG) are spelled out upon first occurrence in the main text.

**Response:**

We have checked the manuscript for acronyms and ensured that they are spelled out upon first occurrence in the main text. In particular, we expanded scientific machine learning (SciML), feature-wise linear modulation (FiLM),

parameter-efficient fine-tuning (PEFT), squeeze-and-excitation (SE), and attention gates (AGs) at their first occurrences.

**Comment 12**

The description of the connection between CNNs and GNNs in Section 2.1, while mathematically rigorous, is somewhat verbose. Consider condensing the high-level analogy to improve the pacing of the methodology section.

**Response:**

We have condensed the high-level CNN–GNN analogy in Section 2.1 by removing repeated explanatory text around Fig. 1. The revised text now briefly states the key connection between convolution and message passing, while the detailed mathematical explanation remains in Appendix A.

**Comment 13**

The manuscript would benefit from a schematic diagram in the Introduction illustrating the overall multi-fidelity transfer learning workflow. This would enhance the structural clarity for the reader.

**Response:**

While a schematic figure could be useful, the study combines two architectures, two fidelity levels, different data representations, several transfer-learning strategies, and two evaluation grids. A single compact schematic would therefore risk becoming overly complex or oversimplifying the workflow. Instead, we have strengthened the textual signposting in the Introduction to more clearly summarise the multi-fidelity workflow.

**Concluding Response.** We again thank the reviewer for the careful reading. We believe the revisions improve the clarity of the manuscript and better communicate the strengths, limitations, and practical implications of the ARU-Net and GNO approaches.

# Authors' Response to Reviewer 2

**General Comments.** The manuscript entitled "Convolutional versus graph-based surrogate models for inter-farm wake prediction using multi-fidelity transfer learning" presents a systematic comparison of two neural-network surrogate models for inter-farm wake deficit prediction, i.e. an attention residual U-Net (ARU-Net) and a graph neural operator (GNO), both pre-trained on low-fidelity TurbOPark samples and fine-tuned on high-fidelity RANS-AWF samples. Averaged over the wake region, the GNO achieves a lower RMSE (0.024 vs. 0.028 m/s) while the ARU-Net attains a higher F1 score (0.98 vs. 0.91), and transfer learning is shown to benefit the ARU-Net substantially more than the GNO. The paper is a useful and timely contribution to the wind farm modelling literature.
The paper is clearly written, well structured, and it is enjoyable to read. The comments below can be used to improve the quantitative claims and fix a few production issues of the manuscript.

**Response:**

Thank you for your positive review, we agree with your assessment and have strived to accommodate the suggested changes as much as possible.

### Main comments

**Comment 1**

Throughout the paper and in the abstract, the RMSE (0.024 vs. 0.028 m/s) is reported, which does not clearly show how good the model is for the readers. Both models appear to be very accurate just from these values. A more meaningful relative error could be designed to better showcase the model performance over analytical models. For example, as a suggestion, this may be defined as the RMSE of ML models normalized by RMSE of TurbOPark model. A key question that should be answered will be, by what extent the developed ML models are better than engineering wake models? This will be key to demonstrating the value of this work.

**Response:**

Although the formulations behind TurbOPark and RANS-AWF differ substantially, the transfer learning results, particularly for the ARU-Net, indicate that TurbOPark reproduces several features also present in RANS-AWF, such as stronger wake decay at higher wind speeds and turbulence intensities, and stronger wakes downstream of denser regions of the wind farm. That said, TurbOPark was not designed to reproduce RANS-AWF data, so reporting the metrics relative to an analytical wake model could be perceived as somewhat misleading and would place the model in an unfairly negative light. We therefore chose not to include this comparison in the main results, but to accommodate readers who would nonetheless like to see it, we have added it in Tab. 5. See also our replies to Comments 5 and 6, which address these changes more directly.

**Comment 2**

The PLayGEN-generated layouts seem pretty realistic and meaningful for real-world applications. However, for evaluations, are all high-fidelity data samples fall into the Cluster type? If it is the case, it will be beneficial to add new test results for samples in single string, parallel string, multiple string types, for quantitative and better evaluation of performance and generalizability.

**Response:**

That is correct, the reason as mentioned in the manuscript is that the RANS-AWF dataset was repurposed from previous work, and as such, it is not easily extendable. As you also note in comment 3, this is not ideal, and we have expanded on the implications this entails and how to solve this in the reply to comment 3.

**Comment 3**

The author clearly described the reason behind using different layout generation methods in low-fidelity and high-fidelity datasets. Still, it would be better if the authors can comment on the pros/cons of PLayGen vs random turbine removal.

**Response:**

The layouts employed correspond to the type 'cluster' in PlayGEN; the remainder of the layout types in PlayGEN are essentially absent from the dataset. This point has been clarified through the following addition to the text:

*As the RANS-AWF dataset used in this work only contains the equivalent of the cluster layout type, the resulting models are therefore not expected to generalize as well to the remaining layout types. Future datasets should incorporate a broader range of layouts to address this limitation in subsequent works.*

**Comment 4**

The impact of this work may be enhanced by open-sourcing the dataset (at least the 200 test cases used in the paper). This is not mandatory though. But it could serve as useful benchmark for the communities in the future.

**Response:**

We have now made the dataset publicly available.

**Comment 5**

Figure 11 could benefit from adding the analytical model results as well, to illustrate how and where the ML models perform better.

**Response:**

We thank the reviewer for this suggestion. We have explored adding an analytical baseline, but as shown in Fig. 1, the TurbOPark profiles deviate substantially from the RANS-AWF reference and the ML predictions. We note that TurbOPark was not formulated to reproduce RANS-AWF data, so a direct comparison on these profiles risks being misleading rather than illustrative. We have therefore chosen not to include it in Figure 11.

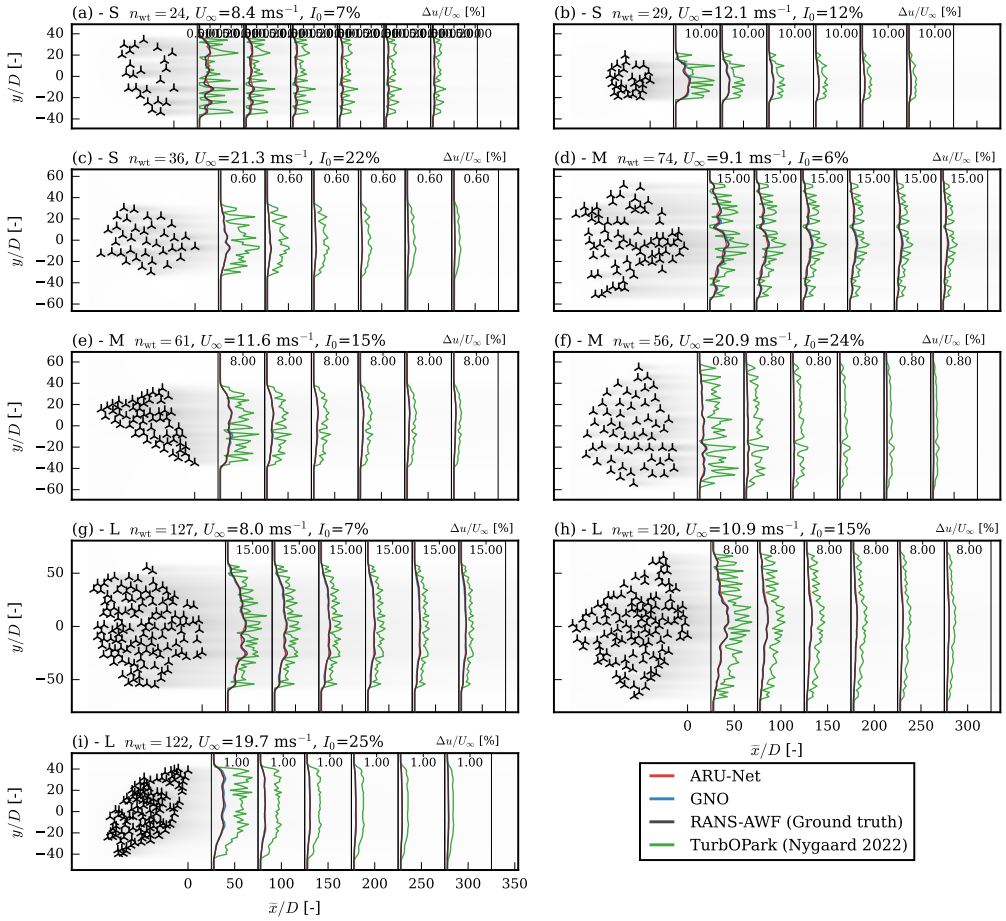


Figure 1: Figure 11 with the TurbOPark baseline added for reference.

**Comment 6**

A general aspect: how the reported evaluation and metrics are affected by intra-farm wakes? Are they minor in the reported RMSE and F1 scores?

**Response:**

We have now added a baseline with respect to a TurbOPark baseline to add this context. Again it should be stated that TurbOPark was not calibrated with RANS-AWF in mind and these results cannot be used to pass judgement on TurbOPark. We have also added a bit of discussion about the F1 score.

### Minor points

**Comment 7**

In abstract, "First, pre-trained on low-fidelity engineering model simulations, and second fine-tuned on high-fidelity Reynolds-averaged Navier-Stokes actuator wind farm (RANS-AWF) data." This is not a sentence.

**Response:**

The incomplete sentence has been restructured to be included in the previous sentence: *Both architectures are trained using multi-fidelity transfer learning, pre-training with low-fidelity engineering model simulations, and then fine-tuning with high-fidelity Reynolds-averaged Navier-Stokes actuator wind farm (RANS-AWF) data.*

**Comment 8**

In abstract, "Transfer learning substantially benefits the ARU-Net, while the GNO shows only marginal improvement." The second half of the sentence is misleading. It should be "transfer learning just improves the GNO marginally".

**Response:**

The sentence in the abstract has been reworded.

**Comment 9**

In Introduction, Line 80-85, CNN-based models based on multi-fidelity data is briefly reviewed. Please note the omission of a related work "Multi-fidelity modeling of wind farm wakes based on a novel super-fidelity network", where a different strategy is used for CNN-based wake modelling based on multi-fidelity data.

**Response:**

We have now made reference to the missing piece of litterature.

**Comment 10**

Line 95-100, “pretraining” should be “pretrained”; “fine-tuning” should be “fine-tuned”.

**Response:**

These have been corrected

**Comment 11**

Why CNN grid ($256 \times 128$) and GNN grid ($257 \times 129$) are not exactly the same but have a difference of 1?

**Response:**

The difference was not intentional; it arose from a miscommunication between the two main authors and did not become apparent until the final stages of the project. The ARU-Net requires a computational grid with side lengths that are powers of two to operate efficiently. The grid was constructed so that the number of cells met this requirement, which in turn meant that the number of nodes did not, since a grid with N cells per dimension has N+1 nodes. As the RANS-AWF data are defined on the nodes, we therefore interpolated the nodal (corner) values to the cell centres to obtain a field on a suitable grid for the ARU-Net. The GNO, by contrast, can operate on an arbitrary grid, so the discrepancy went unnoticed until we compared the results of the two models. Rather than retrain both models, we chose to report both sets of numbers. This also allowed us to highlight a qualitative difference between the two approaches, namely the grid independence of the GNO.

**Comment 12**

Figure 3 appears later in the main texts than Figure 2. Please order the figures by their order of appearance.

**Response:**

We believe the reviewer has missed the reference made to Fig. 2 on page 6 L148, as it is squeezed in between Fig. 1 and Fig. 2. The reference should not be this easy to overlook, but should not be an issue in the final version for publication.

**Comment 13**

Figure 13 has no green color. But the main texts refer to the green color. Please fix. Also check other figures and the related texts.

**Response:**

This was meant to be the color red, this has now been corrected.

**Comment 14**

Figure 12: (a-e) labels do not match the ones in the figure.

**Response:**

The rows and columns had erroneously been switched, this has been corrected in the figure now.

**Comment 15**

Broken cross-reference and links: line 562, "As mentioned in ??"; "...made available at Zenodo (?)". Please fix.

**Response:**

The first reference has been removed and the link to the dataset on Zenodo has been added.

**Comment 16**

Line 600: "Regarding fine-tuning the full-tuning and frozen encoder strategy performed comparably for the ARU-Net" reads awkwardly; please rephrase.

**Response:**

The paragraph has been rephrased:

*For the ARU-Net, full fine-tuning and the frozen-encoder strategy performed comparably. Neither the ARU-Net nor the GNO improved from LoRA-based PEFT, which offered no significant computational or accuracy benefit over training from scratch, likely due to the relatively small size of the models compared to the large language models for which LoRA was originally designed.*

**Concluding Response.** Once again we would like to thank you for your constructive review of our manuscript. We believe the updates has made the manuscript stronger and believe it now meets the quality necessary for publication.