Editor's and reviewers' comments appear in italics; **our responses appear in boldface blue text**.

Associate Editor

Both reviewers agree that you have incorporated most of the suggestions very well and that the article is much more to the point now and less diluted across the different sensitivity studies. However, Reviewer 2 remains concerned about the limited temporal range of the measurements and model simulations. I agree with this assessment and recommend a second round of major revisions.

We thank the editor for their thoughtful review of the manuscript as well as the comments from both reviewers. We have included an extended evaluation at the FINO1 tower for the month of the October 2017 in the revised version of this manuscript. (Due to the end of our project, we do not have access to computational resources that would allow us to simulate a longer time period than what is presented here.) We believe that this additional assessment which we provide here, as well as the itemized edits in the reviewer responses below, should address any outstanding concerns.

In Appendix A4, starting on line 616:

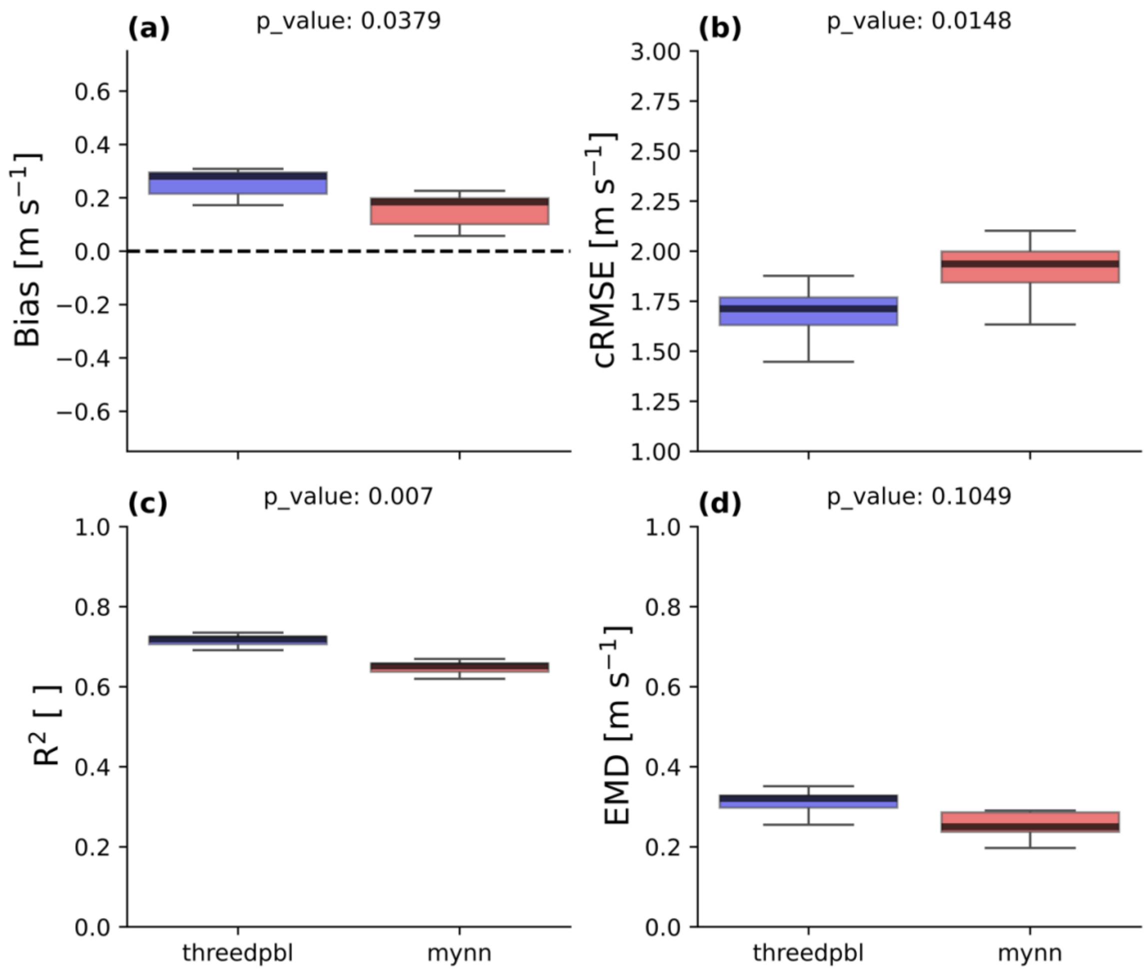

Figure A7. Error metrics across all Fitch simulations with the advection option on and with a wind farm TKE factor of 0.25 at the FINO1 site for the period of Oct 1 - Oct 28 for wind speed. The bar shows the median. The box encloses the interquartile range (IQR), and the whiskers extend to Q1-1.5*IQR and Q3+1.5*IQR. (a) wind speed bias; (b) wind speed cRMSE; (c) wind speed R2; (d) wind speed EMD.

"Here we provide an extended comparison of the performance of the two PBL schemes by comparison to observations at the FINO1 site. This evaluation was designed to reflect the FINO1 evaluation performed in the main manuscript as closely as possible. As such, wind speed data were analyzed at the same seven (34 m, 41 m, 51 m, 61 m, 81 m, 91 m, 102 m) heights at the same 10 min temporal resolution considered in the main manuscript. The evaluation was also performed with the same model cell with the same four error metrics (bias, cRMSE, R2, and

EMD) and Mann–Whitney U statistical significance testing. The evaluation period is the key difference between the main FINO1 evaluation and that performed here. The evaluation period (1 October - 28 October) contains all the contiguous, numerically-stable days in October 2017. Note that to ensure this numerical stability, a timestep of 18 s was selected instead of the 30 s timestep employed in the main text.

"The PBL schemes show mixed performance in the extended FINO1 evaluation. 3DPBL produces lower median cRMSE (Fig. A7b) and higher R2 (Fig. A7c) than MYNN, whereas MYNN produces lower bias and lower EMD (Fig. A7d) than 3DPBL. However, only the differences in cRMSE, bias, and R2 are statistically significant; the EMD (Fig. A7d) difference is not. Considering only statistically significant error metrics, 3DPBL performs better for two of the three metrics.

"The longer time period also clarifies model differences. In the twelve hour analysis, statistically significant differences only emerged for cRMSE (Fig. 14d), whereas the 28 day analysis instead suggests statistically significant differences with respect to three error metrics (Fig. A7). In both cases, 3DPBL shows the optimal cRMSE. Thus, increasing the length of the evaluation window helps reveal systematic differences between the models."

Reviewer 1

Second round review of "A North Sea in situ evaluation of the Fitch Wind Farm Parametrization within the Mellor–Yamada–Nakanishi–Niino and 3D Planetary Boundary Layer schemes" by Agarwal et al. The manuscript has been much improved since the first version, with the authors having made many changes related to the comments of both reviewers. The scope has been narrowed significantly, and results are better connected to theory and existing literature. Still, the analysis presented in the manuscript should be improved before acceptance.

Specific comments

• The manuscript uses two point measurements of very short temporal range to determine what PBL scheme is more accurate. The actual WS values at these points are a resultant of many factors, mainly the background wind speed, wake generation and wake recovery. These factors should be addressed separately and in more detail.

From my understanding:

o Background wind speed: higher for MYNN

o Generated wake: higher for MYNN due to higher wind speed and thus higher trust. Otherwise the same (Fitch code not touched).

o 3DPBL adds more TKE, especially around upper tip height This leads to weaker wakes in 3DPBL: Lower trust due to lower WS, and faster wake recovery due to more turbulence.

We appreciate the opportunity to elaborate on the mechanisms for these differences. While the three bullet points that the reviewer indicates are true, we believe that the more physically intuitive way to describe the relationships between these factors is by starting with the differences in TKE between the PBL schemes. As such, we have elaborated upon and re-written the "Spatial Differences" section to hopefully make this clearer. Starting on line 355:

"Wind field behavior near the turbines differs from that for the rest of the simulation domain, on average. Recall that the 3DPBL scheme incorporates more potential sources of TKE generation. MYNN average wind speeds are faster than 3DPBL average wind speeds outside of the turbine wakes (Fig. 8a). These MYNN average wind speeds likely exceed 3DPBL average wind speeds in this area because the 3DPBL scheme generates more TKE (Fig. 8b) due to the fact that the 3DPBL turbulence parameterization includes more terms that allow for horizontal mixing. Because the atmospheric boundary layer experiences more TKE in the 3PBL simulation, the frictional forces are slightly larger, reducing the horizontal winds in this stably stratified flow case. This larger TKE with the 3DPBL scheme extracts more momentum from the mean wind, resulting in a greater reduction in wind speeds. This finding, that MYNN wind speeds are faster than 3DPBL wind speeds for the same forcing or boundary conditions, is consistent with other comparisons of these two PBL schemes, completed in both real and idealized conditions (Juliano et al., 2022; Rybchuk et al., 2022; Arthur et al., 2022; Peña et al., 2023; Arthur et al., 2024). This finding is also documented further in Fig. 7 and in Sect. A1. Note that this increased TKE in the 3DPBL scheme is particularly evident along the Danish coast, resulting in even slower winds in the 3DPBL scheme (Fig 8b).

"Once these faster winds, in the MYNN scheme, enter the region of a wind farm, the faster winds lead to a larger drag force exerted by the wind turbines (Fig 3b) and therefore a larger wake effect, resulting in slower winds in the MYNN wakes (Fig. 8a). This distinct behavior in the wakes arises from differences in the drag forces for each PBL scheme. The drag forces are very sensitive to wind speed (Fig 3b). Because the MYNN wind speeds are slightly faster when entering the

wind farms, the resulting MYNN drag force (Eq. 15, Fig 3b) is generally larger than the 3DPBL drag force. As a consequence, the MYNN scheme shows stronger and longer wakes than the 3DPBL scheme, on average (Fig. 8c). The MYNN average wind speed reduction is sufficiently strong such that 3DPBL average wind speeds exceed MYNN average wind speeds within the turbine wake (Fig. 8a). Further, because 3DPBL average wind speeds exceed MYNN average wind speeds in this region, the 3DPBL scheme also has more turbine-induced TKE than the MYNN scheme (Fig. 8b). This turbine-induced TKE can help erode the wake in the 3DPBL simulations.

"This behavior emerges fundamentally because of the difference in ambient background turbulence between the 3DPBL scheme and the MYNN scheme in these stably stratified conditions. Other behavior would manifest in stably stratified conditions if winds exceeded the wind speed of the peak drag force (near 13 m/s in Fig 3b), such that faster winds would result in a smaller drag force."

We also expand upon this discussion in Appendix A1 starting on line 530:

"MYNN average wind speeds (Fig. A1c,d) are also consistently faster than 3DPBL average wind speeds (Fig. A1a,b) for both NWF and Fitch cases outside of the turbine wakes (Fig. A1e,f). This difference in wind speeds between the PBL schemes can be explained by TKE differences between the two PBL schemes. Because the 3DPBL scheme has larger TKE (Fig. 8b), the 3DPBL scheme extracts more momentum and reduces wind speeds further. In contrast, Fitch and NWF wind speeds differ near the turbines. Notably, 3DPBL Fitch average wind speeds exceed MYNN Fitch average wind speeds in the turbine wakes (Fig. A1f). This reversal of which PBL scheme shows the faster average wind speed can be explained by differences in the turbine drag force between the two PBL schemes. The MYNN scheme has a stronger turbine drag force (Eq. 11, Fig. 3b) because of its faster initial wind speeds, which also implies that the MYNN scheme has stronger and deeper wakes (Fig. 8c). This reversal (of which PBL scheme shows the higher average wind speed) occurs only in the monotonically increasing region of the drag (proxy) curve (Fig. 3b). If the wind speeds were instead within the monotonically decreasing region of the drag (proxy) curve (Fig. 3b), MYNN wind speeds would likely exceed 3DPBL wind speeds even in the wakes because the faster winds would result in a smaller drag force (Fig. 3b). Given that NWF wind speeds mirror Fitch wind speeds outside of the turbine wakes and NWF wind speeds differ from Fitch wind speeds within the wakes, the dominant

mechanism for these differences is more likely related to the turbines and not to the underlying meteorology."

The point measurement are insufficient to capture these dynamics and should be evaluated as such. It should be clearly indicated that you use one point in a wind farm within the rotor area (FINO1) and one point above the wind farm (flights).

We appreciate the reviewer's concern over the spatial representativeness of the measurements. We emphasize these site differences first and foremost in the abstract: "the FINO1 site located at a single point within the turbine rotor layer" vs. "aircraft observations taken 100 m above the wind farm." We also reinforce this distinction in our Section 4 Model Evaluation when we reference "observations collected 250 m above the surface and 100 m above a wind farm." This distinction is perhaps emphasized when we highlight how differences in measurement heights dictate performance differences. This discussion begins on line 473:

"Differences in the relative measurement height between the two sites in this case study may also affect the PBL comparison. Whereas the FINO1 tower is within the turbine rotor region, the aircraft measurements are taken more than 100 m above the turbines in a stably stratified boundary layer (Fig. 4) that suppresses some interactions between the atmosphere sampled at the turbine level and the aircraft level (Fig. 6). Thus, the 3DPBL scheme improves cRMSE turbine-induced turbulence characterization in the turbine rotor region as sampled at FINO1 (Fig. 14d) and overpredicts turbine-induced turbulence aloft as sampled by the aircraft measurements, on average (Fig. 14c)."

To that end, we also wish to clarify that the aircraft measurements are not in fact point measurements. While they are constrained in altitude, they reflect broader spatial variability.

More trajectories inside and downstream of the wind farm are needed to fully evaluate what PBL scheme captures the wake dynamics more accurately. While ideally a recommendation is given on what scheme to use, this is not possible based on the current analysis and observations used. This should be noted more clearly, especially in the abstract.

We thank the reviewer for this comment. We agree that future evaluations could leverage newly-available observations to determine more information about physical differences between these schemes. To reflect this sentiment, we add a

line at the end of the abstract: "Future evaluations across broader temporal and spatial scales may offer further insight into model differences."

This claim is then complemented by a suggestion for future work in the conclusion. Starting on line 505:

"Subsequent investigations could explore other case studies to provide perspective into the generalizability of the results across other sites. Similarly, datasets from the third Wind Forecast Improvement Project (WFIP3) could be useful to explore how offshore wind characterization might differ between the North Sea and the eastern United States (WFIP3). Moreover, datasets from the land-based, horizontally homogeneous American WAKE experimeNt – or AWAKEN – campaign (Moriarty et al., 2024) could be useful to study because previous land-based studies analyzing the 3DPBL scheme have involved complex terrain and far fewer detailed observations."

• *The FINO1 dataset allows for a comparison over a longer temporal range. While I appreciate the desire for consistency with other datasets and literature, I recommend a longer simulation (at least a month, ideally a year) to evaluate whether the results presented in the presented case study hold over a longer period of time.*

We thank the reviewer for an opportunity to extend the timeframe of our FINO1 analysis. Here we provide an extended evaluation at the FINO1 site for the month of October 2017 in Appendix A4. Unfortunately, longer simulations are simply not possible given the limited computational resources now available to us.

Starting on line 616:

"Here we provide an extended comparison of the performance of the two PBL schemes by comparison to observations at the FINO1 site. This evaluation was designed to reflect the FINO1 evaluation performed in the main manuscript as closely as possible. As such, wind speed data were analyzed at the same seven (34 m, 41 m, 51 m, 61 m, 81 m, 91 m, 102 m) heights at the same 10 min temporal resolution considered in the main manuscript. The evaluation was also performed with the same model cell with the same four error metrics (bias, cRMSE, R2, and EMD) and Mann–Whitney U statistical significance testing. The evaluation period is the key difference between the main FINO1 evaluation and that performed here. The evaluation period (1 October - 28 October) contains all the contiguous, numerically-stable days in October 2017. Note that to ensure this numerical stability, a timestep of 18 s was selected instead of the 30 s timestep employed in the main text.

"The PBL schemes show mixed performance in the extended FINO1 evaluation. 3DPBL produces lower median cRMSE (Fig. A7b) and higher R2 (Fig. A7c) than MYNN, whereas MYNN produces lower bias and lower EMD (Fig. A7d) than 3DPBL. However, only the differences in cRMSE, bias, and R2 are statistically significant; the EMD (Fig. A7d) difference is not. Considering only statistically significant error metrics, 3DPBL performs better for two of the three metrics.

"The longer time period also clarifies model differences. In the twelve hour analysis, statistically significant differences only emerged for cRMSE (Fig. 14d), whereas the 28 day analysis instead suggests statistically significant differences with respect to three error metrics (Fig. A7). In both cases, 3DPBL shows the optimal cRMSE. Thus, increasing the length of the evaluation window helps reveal systematic differences between the models."

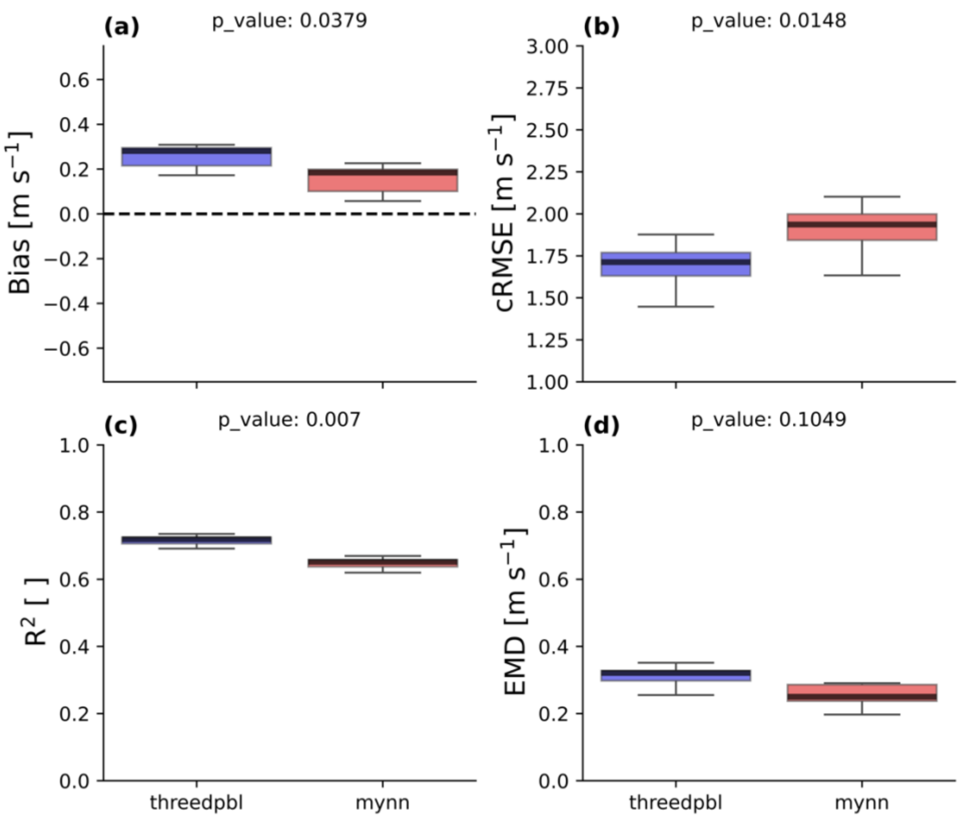

Figure A7. Error metrics across all Fitch simulations with the advection option on and with a wind farm TKE factor of 0.25 at the FINO1 site for the period of Oct 1 - Oct 28 for wind speed. The bar shows the median. The box encloses the interquartile range (IQR), and the whiskers extend to Q1-1.5*IQR and Q3+1.5*IQR. (a) wind speed bias; (b) wind speed cRMSE; (c) wind speed R2; (d) wind speed EMD.

Technical corrections

• Introduction: Many references to the same papers in a row. This should be rewritten to improve readability.

We appreciate the reviewer's feedback on readability. We agree that individual sentences that include multiple references to the same paper can be structured more clearly. We have thinned the North Sea references section starting on line 50 significantly.

"A recent North Sea measurement campaign has stimulated interest in WFP intercomparison and validation efforts. The Wind Park Far Field (WIPAFF) project was an aircraft expedition to understand offshore wind wake behavior in the German Bight. This expedition took place in a location with multiple wind farms, with 41 total flights between 6 September 2016 and 15 October 2017 (Platis et al., 2018; Bärfuss et al., 2021). Siedersleben et al. (2018b) modeled a case study that included one day of these aircraft observations and found that, because of challenges in characterizing air-sea interactions with a mesoscale model, improving background flow characterization contributed greater model improvement than any wind farm-specific parameter configuration. Siedersleben et al. (2018a) then extended the analysis of this one case study to demonstrate the presence of hub-height potential temperature and water vapor wakes during stably-stratified conditions. Siedersleben et al. (2020) then leveraged 3 days of these aircraft observations that occurred during stable conditions to explore the sensitivity of grid cell spacing and TKE advection within the Fitch parameterization. Larsén and Fischereit (2021) extended the work of Siedersleben et al. (2020) by comparing the explicit wake parameterization (EWP) (Volker et al., 2015) and Fitch (Fitch et al., 2012) schemes and exploring model performance during wind farm interactions with low-level jets (LLJs) and this work was then extended to introduce wave coupling (Larsén et al. (2024). Ali et al. (2023) also used data from one day of the Siedersleben et al. (2020) case study to validate five (Fitch et al., 2012; Volker et al., 2015; Abkar and Porté-Agel, 2015; Pan and Archer, 2018; Redfern et al., 2019) common WFPs with aircraft measurements as well as nearby meteorological tower and synthetic aperture radar observations."

• L144: justify the use of the boundary layer approximation

We thank the reviewer for this curiosity. We clarify that this decision is consistent with previous analyses and therefore facilitates comparisons with the previous literature. Starting on line 125:

“A ‘boundary layer’ approximation of the 3DPBL scheme retains the same 3DPBL length scale formulation (and corresponding closure constants) but instead sets the horizontal gradients of the mean quantities and the vertical gradient of the vertical velocity to zero. This 3DPBL approximation is the most commonly-used version of the 3DPBL scheme (Rybchuk et al., 2022; Arthur et al., 2022; Sanchez Gomez et al., 2023; Peña et al., 2023; Arthur et al., 2024), and we employ it in this analysis for consistency.”

• *Section 2.3: This section is messy and should be rewritten. Information related to sensitivity tests included in the Appendices should be moved there.*

We thank the reviewer for an opportunity to be clearer in our descriptions. We have moved all sensitivity discussions to their respective Appendices. We also removed the theoretical walkthrough of the Fitch scheme. We have also removed several potentially-redundant sentences. Starting on line 173:

“Simulations were performed with the WRF model (Skamarock et al., 2021) with the Fitch WFP (Fitch et al., 2012), modified to incorporate the 3DPBL scheme. The WRF simulations here generally followed the setup of Ali et al. (2023) (Table 5). Simulations represented the single day of 14 October 2017 with a 30 s timestep in the outer domain and a 10 min output, starting at 00:00:00 UTC with a 12 hour spin-up period so that the analysis period starts at 12:00:00 UTC. The region was simulated using three nested domains with an outer horizontal grid size of 15 km and a nesting ratio of 3 so that the innermost domain has a grid size of 1.67 km. Eighty vertical levels were employed with 17 levels were lower than 200 m, and between 8 and 12 levels intersected the turbine’s rotor. We used ERA5 reanalysis (Hersbach et al., 2020) for initial and boundary conditions, the WRF double-moment six-class microphysics scheme (Hong et al., 2010), the RRTMG shortwave and longwave radiation scheme (Mlawer et al., 1997), the Noah land-surface model (Niu et al., 2011) and the Kain–Fritsch cumulus parameterization scheme (Kain, 2004) in the outer domain only.

“Differences also exist between the WRF setup presented in this work and that used in Ali et al. (2023). Here, we varied the PBL scheme to explicitly consider the influence of the PBL scheme on wake behavior for the Fitch WFP. Two PBL schemes were considered: level 2.5 MYNN ("MYNN") and the NCAR 3DPBL scheme with the PBL approximation ("3DPBL") as described in Rybchuk et al. (2022). The MYNN scheme is activated in all outer domains for all simulations. Further, Ali et al. (2023) performed their analysis on a modification of WRF v4.5.1,

whereas the analysis presented in this work relied on an earlier version of WRF (V4.4.2) in which the 3DPBL scheme is integrated.

“Wind farm effects were represented with the Fitch WFP. The wind-speed- and turbine-model-dependent thrust and power coefficients were integrated into the WRF model through turbine specification files (Fig. 3). Individual turbines were also integrated into the WRF grid with a file from Ali et al. (2023) that contains a given turbine’s latitude, longitude, and turbine type. We used files from the Ali et al. (2023) repository and extracted the key information to fit the standard Fitch WFP format.

“Our simulations varied the wind farm option, wind farm TKE factor, and the advection option for each PBL scheme (Table 6). We highlight the results from the two simulations with the Fitch scheme, the advection option on, and the wind farm TKE factor of 0.25 to focus the scope on the effects of the PBL scheme (Table 6). The results from the other fourteen runs are analyzed throughout the Appendix.”

• *Section 3.1.1: This section is very long and hard to read. Please split up in more subsection and rewrite to improve readability.*

We thank the reviewer for an opportunity to improve the readability of our Atmospheric stability section. This section has now been split into three subsections, separating out “Aircraft profile flights” and “Impacts of stable stratification”. Starting on line 287:

The modeled atmospheric stability at both the FINO1 location (Fig. 4a,b) and over the aircraft region (Fig. 4f,g) suggests a weakly stable profile near the surface with stronger stability aloft. MYNN simulations are slightly warmer (according to potential temperature) and more stable near the surface than 3DPBL simulations (Fig. 4b,g). For both sites, these modeled temperature profiles are consistent with available observations (Fig. 4a,f) and the observed potential temperature in the aircraft region is slightly warmer than the modeled potential temperature (Fig. 4g). Because the modeled air temperatures are almost identical between models (Fig. 4a,f), these slight differences in the surface stability could be a consequence of the greater TKE with the 3DPBL scheme that encourages slightly more mixing. The wind direction profiles at both FINO1 (Fig. 4c) and the aircraft transect (Fig. 4h) regions show veering wind (i.e., the wind direction rotates from southwesterly direction near the surface to westerly aloft), suggesting warm-air advection. Simulations with MYNN also tend to have slightly more southerly295

winds than simulations with the 3DPBL scheme (Fig. 4c,h). Wind speeds are similar near the surface and through the rotor layer with slight variations between the aircraft and FINO1 regions (Fig. 4d,i). Peak wind speeds for both sites are slightly higher than the peak TKE, reflecting a LLJ, as in Larsén and Fischereit (2021). 4j). The wind direction vertical profiles for both the FINO1 (Fig. 4c) and aircraft (Fig. 4h) regions also suggest inversions at 500 m. Although this separate air mass would not be considered an inversion according to the potential temperature vertical profiles (Fig. 4a,d), the variation in wind speed does support a distinct layer above 500 m (Fig. 4d,i). Further, the modeled PBL height, which in stably stratified conditions is determined as the height at which the TKE profile reaches 5% of its surface value (Olson et al., 2019), corroborates this distinction. Thus, dynamically, the top of the stable boundary layer likely resides around 500 m.

"The TKE peaks differentiate the two model vertical profiles at both sites. 3DPBL TKE is consistently larger beyond the surface and this discrepancy is largest at, and slightly above, the rotor level. The fundamental differences in horizontal mixing and length scales between the schemes lead the 3DPBL scheme to characterize more TKE and these differences are more accentuated both with a LLJ and the turbine-induced TKE contribution. These modeled TKE also differ from observations. While MYNN TKE underpredicts TKE observed during both even and odd transects, the higher 3DPBL TKE overpredicts TKE during even transects and underpredicts TKE during odd transects. These deviations between modeled and observed TKE may, in part, reflect the model formulations. However, yaw (mis)alignment of the aircraft may also play a part in these discrepancies. The error in yaw alignment is expected to have alternating signs on whether the wind approached the sensor from the starboard or the backboard side (Bärfuss et al., 2023). Because the even and odd transects involved opposite alignments of the aircraft, it would follow that the odd and even transects would then show distinct TKE (and wind speed on a smaller scale) errors. Thus, the discrepancy between TKE observed with the odd and even transects may also play a part in the poor model agreement."

Followed by a section on line 316:

3.1.2 Aircraft profile flights

The aircraft profile flights suggest a similar vertical structure to that for the FINO1 and aircraft transect regions. Both MYNN and 3DPBL temperature and potential temperature profiles at the locations of the profiles (Fig. 5a,b) suggest a weakly stable profile near the surface with stronger stability aloft. This stability is reinforced by the aircraft profile observations (Fig. 5f,g). A LLJ again emerges both in wind speed (Fig. 5d,i) and TKE (Fig. 5e,j) maxima. Finally, both modeled (Fig. 5c) and observed (Fig. 5h) wind direction from the aircraft profiles again suggest warm-air advection with an inversion that is supported by the modeled PBL height.

For the aircraft profile flights, the vertical structure suggests slight discrepancies between model and observation both in terms of values and shape (Fig. 5). For example, inversions in temperature (Fig. 5a,f) and potential temperature (Fig. 5b,g) are simulated approximately 100 m higher than observed. The wind direction (Fig. 5c,h) rotation is also modeled at a higher altitude than they are observed. The observed surface wind direction (Fig. 5h) also reflects an approximately 20 degree discrepancy from the modeled surface wind direction (Fig. 5c). Observed wind speeds (Fig. 5i) are also faster than modeled wind speeds (Fig. 5d) with an additional local peak lower than 200 m that is not captured by the model. Finally, observed TKE (Fig. 5j) suggests two peaks, with one peak within the turbine rotor layer and the other peak well above aircraft transect flights and the top of the boundary layer. In contrast, modeled TKE (Fig. 5c) suggests a single TKE peak with a more gradual decrease with height. These discrepancies in agreement could be explained by several factors, such as the larger number of available data points for the aircraft profile flights, a lack of a smoothing time average for the aircraft profile observations, and the spatial variability sampled with the aircraft profile flights.

And a final section on 333:

3.1.3 Impacts of stable stratification

The stable stratification may suppress some of the turbine-generated turbulence from extending aloft. Both FINO1 and the aircraft transect measurement regions show simulated TKE peaks at altitudes within the rotor region due to wind-farm-generated turbulence. Both the FINO1 and aircraft regions also show larger amounts of TKE with the 3DPBL scheme (Fig. 6a,b) than with the MYNN scheme (Fig. 6c,d). These differences in TKE between the PBL schemes, which are consistent with those shown under the idealized, stable conditions simulated in Rybchuk et al. (2022), reflect the fundamental differences between the models. The stronger TKE maxima in the 3DPBL (Fig. 6a,b) at both sites also lead to greater interfarm TKE overlap with 3DPBL than for MYNN (Fig. 6c,d), such that the TKE interactions between the wind farms are more pronounced for 3DPBL. The

difference in both the intensity and degree of overlap between the 3DPBL and MYNN TKE maxima is stronger in the aircraft region (Fig. 6f) than in the FINO1 region (Fig. 6e), likely due to the larger number of wind turbines in the aircraft region (Fig. 6a,c). However, whereas the aircraft region's simulation suggests a higher maximum TKE than the FINO1 region, not all of the aircraft region TKE is captured by the measurements (Fig. 6b,d). The turbine-induced turbulence is sampled well by the FINO1 tower (Fig. 6a,c). However, some of this turbine induced turbulence is suppressed from reaching the aircraft region measurement height (Fig. 6b,d). The two PBL schemes also do not respond to this stable stratification in the same manner. The 3DPBL scheme simulates more of this TKE reaching the aircraft measurement height, resulting in a comparatively larger 3DPBL-MYNN TKE difference at the aircraft measurement height (Fig. 6f).

These diverging patterns in TKE characterization between the two sites also have secondary influences on measured wind speeds (Fig. 7). In the aircraft region, enhanced TKE implies greater momentum extraction and, consequently, reduced wind speeds at the measurement site (Fig. 7b,d,f). In the FINO1 region, however, TKE from aloft mixes more momentum from aloft into the measurement region, which increases wind speeds (Fig. 7a,c,e). Thus, the differing measurement altitudes between the two sites may impact both the TKE and wind speed assessments presented below.

• *Fig 4: the main difference between PBL schemes is clearly in the TKE profiles. This should be highlighted more in the text.*

We thank the reviewer for an opportunity to emphasize one of our key results. We have added two full paragraphs, starting on line 304, that now read:

"The TKE peaks differentiate the two model vertical profiles at both sites. 3DPBL TKE is consistently larger beyond the surface and this discrepancy is largest at, and slightly above, the rotor level. The fundamental differences in horizontal mixing and length scales between the schemes lead the 3DPBL scheme to characterize more TKE. These differences are more accentuated both with a LLJ and the turbine-induced TKE contribution.

Simulated TKE also differs from observations. While MYNN TKE underpredicts TKE observed during both even and odd transects, the 3DPBL TKE overpredicts TKE during even transects and underpredicts TKE during odd transects. These deviations between modeled and observed TKE may, in part, reflect the model formulations. However, yaw (mis)alignment of the aircraft may also play a part in

these discrepancies. The error in yaw alignment is expected to have alternating signs on whether the wind approached the sensor from the starboard or the backboard side (Bärfuss et al., 2023). Because the even and odd transects involved opposite alignments of the aircraft, it would follow that the odd and even transects would then show distinct TKE (and wind speed on a smaller scale) errors. Thus, the discrepancy between TKE observed with the odd and even transects may also play a part in the poor model agreement."

• *Fig 5: please increase quality of the figure*

We thank the reviewer and have increased the quality of this figure by increasing the DPI to 1200.

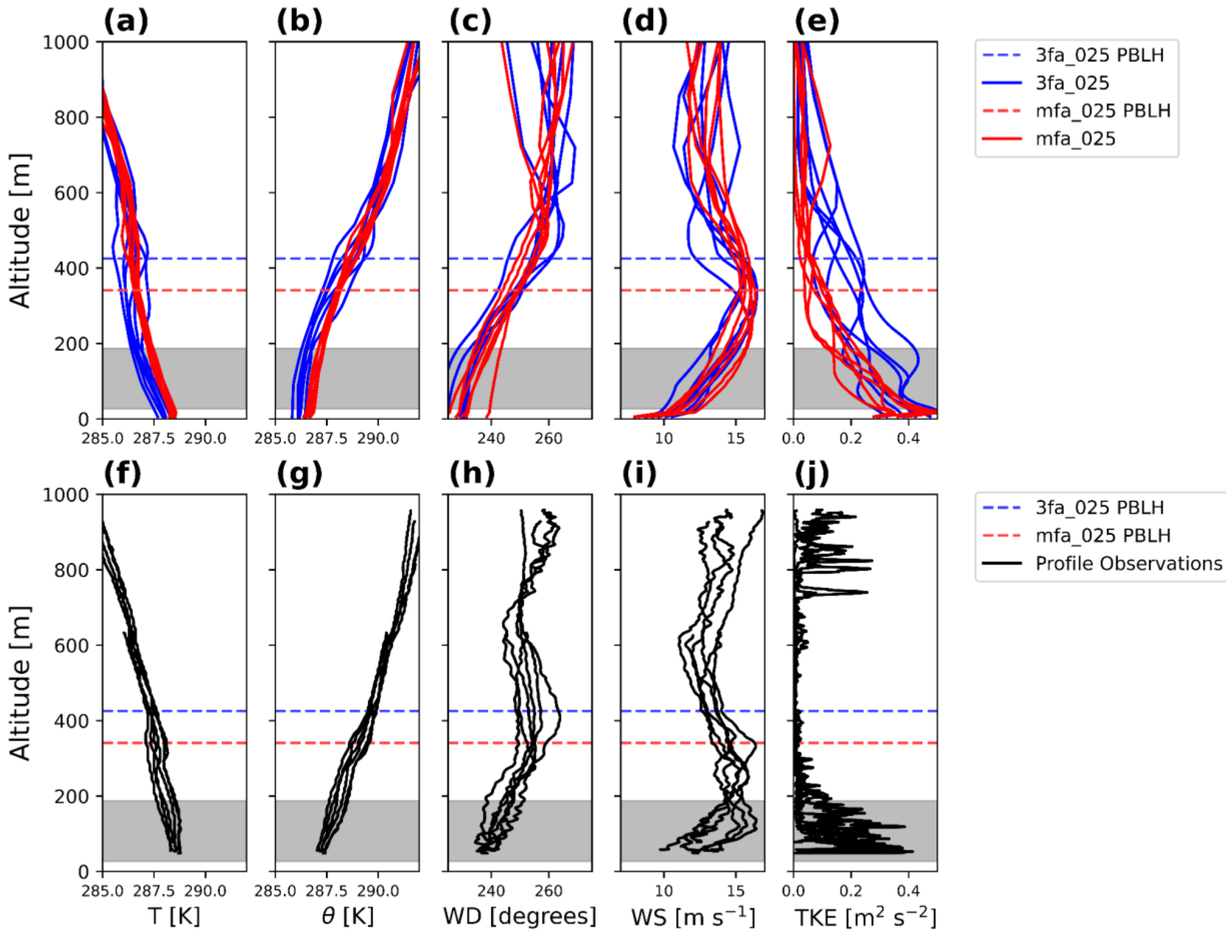

Figure 5. Observed and modeled vertical profiles for the aircraft vertical profile flights (Table 3, Fig. 1b). In all cases, the horizontal line indicates the modeled PBL height, and the color differentiates the PBL scheme. The top row of panels corresponds to modeled output and the bottom row of panels corresponds to the

aircraft profile observations. Modeled output are determined to be a given middle cell for each profile as in Larsén and Fischereit (2021) based on the timestep indicated in Table 3. (a) modeled temperature; (b) modeled potential temperature; (c) modeled wind direction; (d) modeled horizontal wind speed; (e) modeled TKE; (f) observed temperature; (g) observed potential temperature; (h) observed wind direction; (i) observed horizontal wind speed; (j) observed TKE.

- *L375: it is mentioned that the inversions are simulated approximately 100m higher than observed. I don't really see this in the figures. Please clarify.*

We remove this observation from discussion.

- *L385 'minor disagreements': I believe the difference between model and observations are significant and should be acknowledged as such.*

We remove the word "minor."

- *L 408-409 "The 3DPBL … region (Fig. 6a,c).": Rephrase*

We remove this sentence entirely.

- *Fig 7: Reduce the extent of the colorbar in a-d.*

We appreciate the reviewer's suggestion to improve the color bar resolution. We have tightened the color bar resolution to only go to 17.5 m/s.

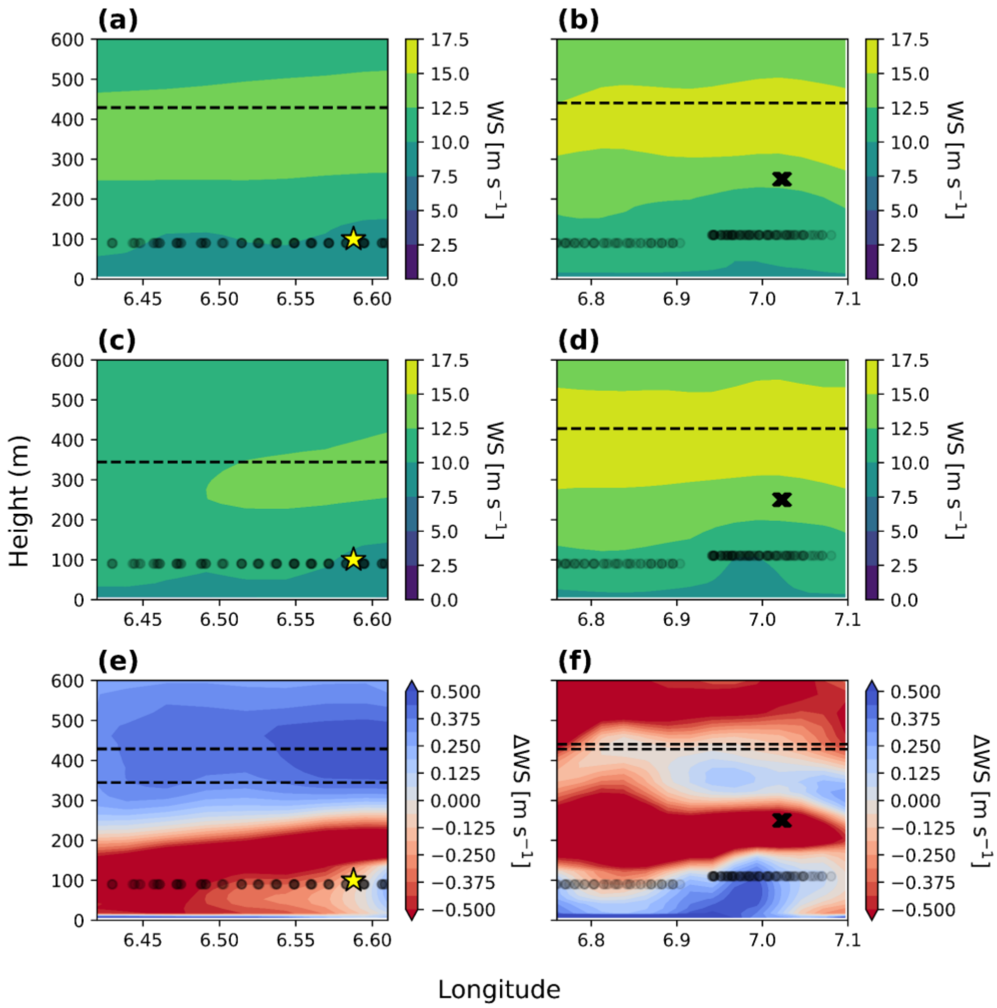

Figure 7. Modeled wind speed cross-section at a constant latitude of 54.03. (a) FINO1 3fa_025; (c) FINO1 mfa_025; (e) FINO1 3fa_025 - mfa_025; (b) aircraft 3fa_025; (d) aircraft mfa_025; (f) aircraft 3fa_025 - mfa_025. The horizontal dashed black line denotes the average modeled PBL height, the star indicates the FINO1 tower location, the "X" marks the first transect path, and the black circles indicate the turbine hub height.

• *L410-411: You claim (throughout the manuscript a few times) that larger TKE is related to greater momentum extraction. Clarify/justify this statement.*

We thank the reviewer for the opportunity to be more explicit and clear in our descriptions. We have written our Spatial variability section to hopefully provide more clarity on the relationship between TKE and momentum extraction. Starting on line 355:

“Wind field behavior near the turbines differs from that for the rest of the simulation domain, on average. Recall that the 3DPBL scheme incorporates more potential sources of TKE generation. MYNN average wind speeds are faster than 3DPBL average wind speeds outside of the turbine wakes (Fig. 8a). These MYNN average wind speeds likely exceed 3DPBL average wind speeds in this area because the 3DPBL scheme generates more TKE (Fig. 8b) due to the fact that the 3DPBL turbulence parameterization includes more terms that allow for horizontal mixing. Because the atmospheric boundary layer experiences more TKE in the 3PBL simulation, the frictional forces are slightly larger, reducing the horizontal winds in this stably stratified flow case. This larger TKE with the 3DPBL scheme extracts more momentum from the mean wind, resulting in a greater reduction in wind speeds. This finding, that MYNN wind speeds are faster than 3DPBL wind speeds for the same forcing or boundary conditions, is consistent with other comparisons of these two PBL schemes, completed in both real and idealized conditions (Juliano et al., 2022; Rybchuk et al., 2022; Arthur et al., 2022; Peña et al., 2023; Arthur et al., 2024). This finding is also documented further in Fig. 7 and in Sect. A1. Note that this increased TKE in the 3DPBL scheme is particularly evident along the Danish coast, resulting in even slower winds in the 3DPBL scheme (Fig 8b).

“Once these faster winds, in the MYNN scheme, enter the region of a wind farm, the faster winds lead to a larger drag force exerted by the wind turbines (Fig 3b) and therefore a larger wake effect, resulting in slower winds in the MYNN wakes (Fig. 8a). This distinct behavior in the wakes arises from differences in the drag forces for each PBL scheme. The drag forces are very sensitive to wind speed (Fig 3b). Because the MYNN wind speeds are slightly faster when entering the wind farms, the resulting MYNN drag force (Eq. 15, Fig 3b) is generally larger than the 3DPBL drag force. As a consequence, the MYNN scheme shows stronger and longer wakes than the 3DPBL scheme, on average (Fig. 8c). The MYNN average wind speed reduction is sufficiently strong such that 3DPBL average wind speeds exceed MYNN average wind speeds within the turbine wake (Fig. 8a). Further, because 3DPBL average wind speeds exceed MYNN average wind speeds in this region, the 3DPBL scheme also has more turbine-induced TKE than the MYNN scheme (Fig. 8b). This turbine-induced TKE can help erode the wake in the 3DPBL simulations.

“This behavior emerges fundamentally because of the difference in ambient background turbulence between the 3DPBL scheme and the MYNN scheme in these stably stratified conditions. Other behavior would manifest in stably stratified conditions if winds exceeded the wind speed of the peak drag force

(near 13 m/s in Fig 3b), such that faster winds would result in a smaller drag force."

We also provide site-specific discussions of the relationship between TKE, momentum extraction, and wind speeds with our vertical slices. Starting on line 349:

"These diverging patterns in TKE characterization between the two sites also have secondary influences on measured wind speeds (Fig. 7). In the aircraft region, enhanced TKE implies greater momentum extraction and, consequently, reduced wind speeds at the measurement site (Fig. 7b,d,f). In the FINO1 region, however, TKE from aloft mixes more momentum from aloft into the measurement region, which increases wind speeds (Fig. 7a,c,e)."

We also justify these explanations by showing how introducing more turbine-induced turbulence reflects these wind speeds patterns for each site in Appendix A2 on line 561:

"As the wind farm TKE factor increases, the 3DPBL wind speed bias decreases (Fig. A3a). Thus, the additional turbulence implies greater momentum extraction in the aircraft region and this greater momentum extraction slows the winds."

And again on line 576:

"For the FINO1 region, both PBL schemes initially underpredict the mean wind speeds and improve the wind speed bias with an increasing wind farm TKE factor (Fig. A4a). In the FINO1 region, increased turbulence in the rotor region extracts more momentum from aloft into the measurement region and increases wind speeds."

- *Fig 8:*

  o What height?

Thank you. The caption for this figure now includes "Difference fields for inner region at a 100 m hub height."

o Blue here indicates 3DPBL>MYNN. In Fig 6e,f and 7e,f blue indicates the opposite. Keep this consistent.

We thank the reviewer for their suggestion. We have swapped the directions of the color bars for both Fig.6 and Fig.7 to allow consistency with Fig. 8.

- *L435 "MYNN wind speeds are also consistently fater than 3DPBL wind speed": I don't quite see this from Fig. 9.*

We remove this observation.

- *L445 “more variable wind speed R2”: I don’t see this from Fig. 10b.*

We remove this observation.

• *Fig 14 and Table 7: combine by annotating values of Table 7 in the subfigures of Fig 14.*

We thank the reviewer for this suggestion. We have added the p values on top of each panel. We also follow the same process for our extended FNO1 evaluation (Appendix A4) and refrain from this practice for the other Appendices where the trend of the simulations are the key insights as opposed to the size of the differences themselves.

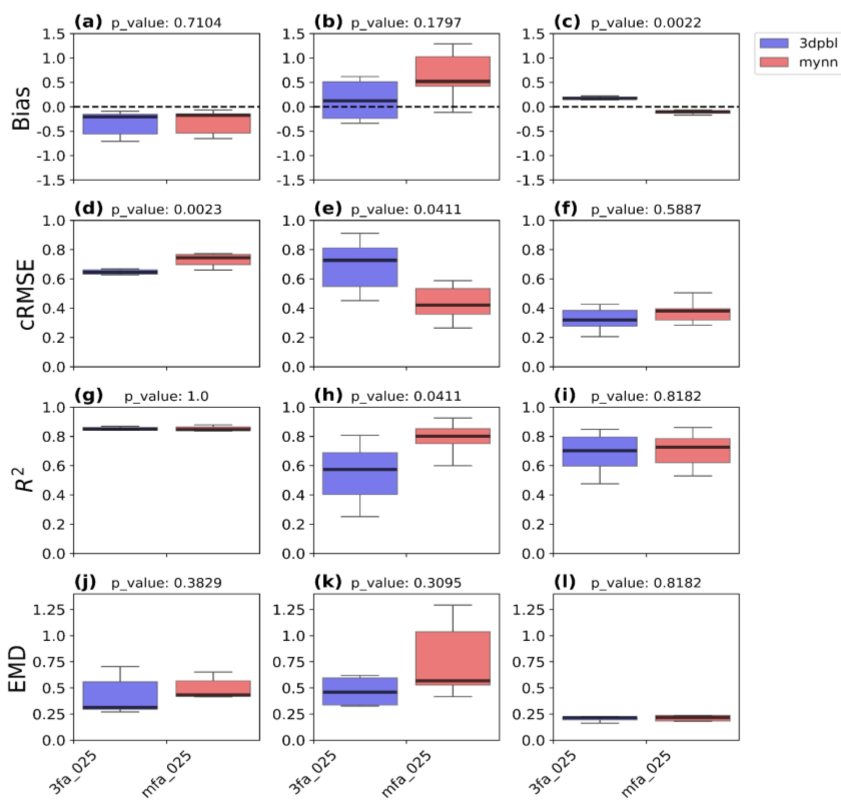

Figure 14. Error metrics across all Fitch simulations with the advection option on and with a wind farm TKE factor of 0.25. The first column corresponds to FINO1 wind speed error metrics, the second column corresponds to aircraft wind speed error metrics, and

the third column corresponds to aircraft TKE error metrics. The bar indicates the median. The box encloses the interquartile range (IQR), and the whiskers extend to Q1-1.5*IQR and Q3+1.5*IQR. Units are non-normalized and reflect the units of the evaluated field. (a) FINO1 wind speed bias [m s−1]; (b) aircraft wind speed bias [m s−1]; (c) aircraft TKE bias [m2 s−2]; (d) FINO1 wind speed cRMSE [m s−1]; (e) aircraft wind speed cRM SE [m s−1]; (f) aircraft TKE cRM SE [m2 s−2]; (g) FINO1 wind speed R2 [ ]; (h) aircraft wind speed R2 [ ]; (i) aircraft TKE R2 [ ]; (j) FINO1 wind speed EMD [m s−1]; (k) aircraft wind speed EM D [m s−1]; (l) aircraft TKE EMD [m2 s−2]).

Reviewer 2

- Line 67: "introducing ocean and wave coupling" should be changed to "introducing wave coupling": there was no ocean coupling done
We thank the reviewer and make this adjustment.

- Eq 1: fj in the equation fk in the description for the Coriolis vector
We thank the reviewer for identifying potentially ambiguous language. fj is a single parameter as part of the full fk Coriolis vector.

- line 396: reference "theory" section on the differences between the schemes
Thank you. We have added a reference citation at this location.

- Eq 20: bias should not be italic, since it is not italic in line 319
Thank you. We have made this adjustment and unitalicized bias.