Response to the reviewer's comments

We thank the reviewer Xin Zhou for this positive and thoughtful review. We provide a detailed response to all comments below with our responses shown in blue. Changes to the manuscript text are included in italic fonts.

This manuscript presents a rigorous, multi-season evaluation of the NOAA HRRR model v4 in the U.S. Northeast coastal marine boundary layer using unique, high-quality remote sensing observations from WFIP3. The work addresses a critical knowledge gap and delivers clear, physically consistent results relevant to offshore wind energy, aviation, and marine forecasting. The paper is well-structured, technically sound, and I have only four major comments:

Major Comments:

1. The manuscript suggests that overestimated SST contributes to warm biases near the bottom of the stability layer and therefore to underestimated static stability. This is plausible and supported by the weak negative correlation between SST error and stability error at NANT and BLOC. However, the correlations are modest, and the relationship is not evident at RHOD. Therefore, the causal wording should be softened in places. For example, instead of implying that SST errors are the primary cause of stability errors, the authors could state that SST errors are "a likely contributing factor," particularly for island sites and southwesterly flow regimes. Other possible contributors, such as boundary-layer mixing, surface-layer parameterization, vertical resolution, land–sea mask representation, and horizontal advection errors, should also be acknowledged. This would make the interpretation more balanced and avoid over-attribution.

We modified our wording, now saying in Sect. 3.3: *This indicates that many of the cases with too-weak static stability in the model were linked to an overestimation of SST and consequently warm biased low-level temperature, suggesting that the specification of the lower boundary conditions are likely a contributing factor to the near-surface model errors.*
In the summary, we highlight now that the correlation between SST and stability errors at the island sites was weak and we make clear that better SSTs could lead to reduced errors of wind and temperature at the island sites, since this is where the weak correlation between SST and static stability errors was found.
We added the following sentence to the summary mentioning other possible factors contributing to model errors: *Other possible contributing factors could be the boundary-layer and surface-layer parameterization, land-sea mask representation, initial and boundary conditions, and errors in horizontal advection.*

2. The authors use a modified Bonner-type LLJ criterion with a wind speed threshold of 8 m s^{-1}, which is lower than in some previous studies. The justification is that many profiles with clear LLJ structure would otherwise be missed. This is reasonable, but the choice of threshold can strongly influence LLJ frequency, CSI, frequency bias, and the classification of weak LLJs. Since about 35–40% of observed LLJs are classified as the weakest class 0, the results may be sensitive to this threshold. I suggest adding a short sensitivity test, perhaps in the supplement,

showing how LLJ frequency and model skill change if a 10 m s^{-1} threshold is used. Even a brief comparison would help readers assess whether the main conclusions are robust to the LLJ definition.

We tested the sensitivity of the model errors and our conclusions to the LLJ criteria by repeating the analysis for LLJ classes 1 through 4 only and produced the respective plots (Figs. 1 to 4 here and Figs. 10 to 13 in the manuscript) using a minimum nose wind speed threshold of 10 m/s. The new plots are included in this response (Fig. 1 to 4). As expected the frequency of observed and simulated LLJ decreased, now ranging between 10 and 13 % in the observations and 8-10 % in the model (Fig. 1), compared to 18-21 % and 14-16 % when class 0 is included. The monthly cycle of average nose height and LLJ frequency as well as the diurnal cycle of LLJ frequency did not change much, except for having slightly lower relative frequencies, as expected. CSI and FB also remained very similar. The HRRR model errors with respect to nose height, wind speed at nose height, shear below and above nose, and stability below and above nose were also very similar (Fig. 2 here and Fig. 11 in the manuscript). The same is true when looking at fixed layer errors (Fig. 3 here and Fig. 12 in the manuscript) or the Bulk Richardson number distribution (Fig. 4 here and Fig. 13 in the manuscript). Because of the overall insensitivity of the errors and conclusions to the chosen nose wind speed threshold, we decided not to include the additional plots using a different nose wind speed threshold in the supplement. Instead we included this additional information in Sect. 2.4: *We tested the sensitivity of our results to the speed threshold by performing the same analysis using a 10 m/s threshold for nose wind speed, that is, only considering LLJ that fall into classes 1 through 4 . While this reduced the number of LLJs in the data set, the model errors were largely the same when removing the weakest LLJ class.*

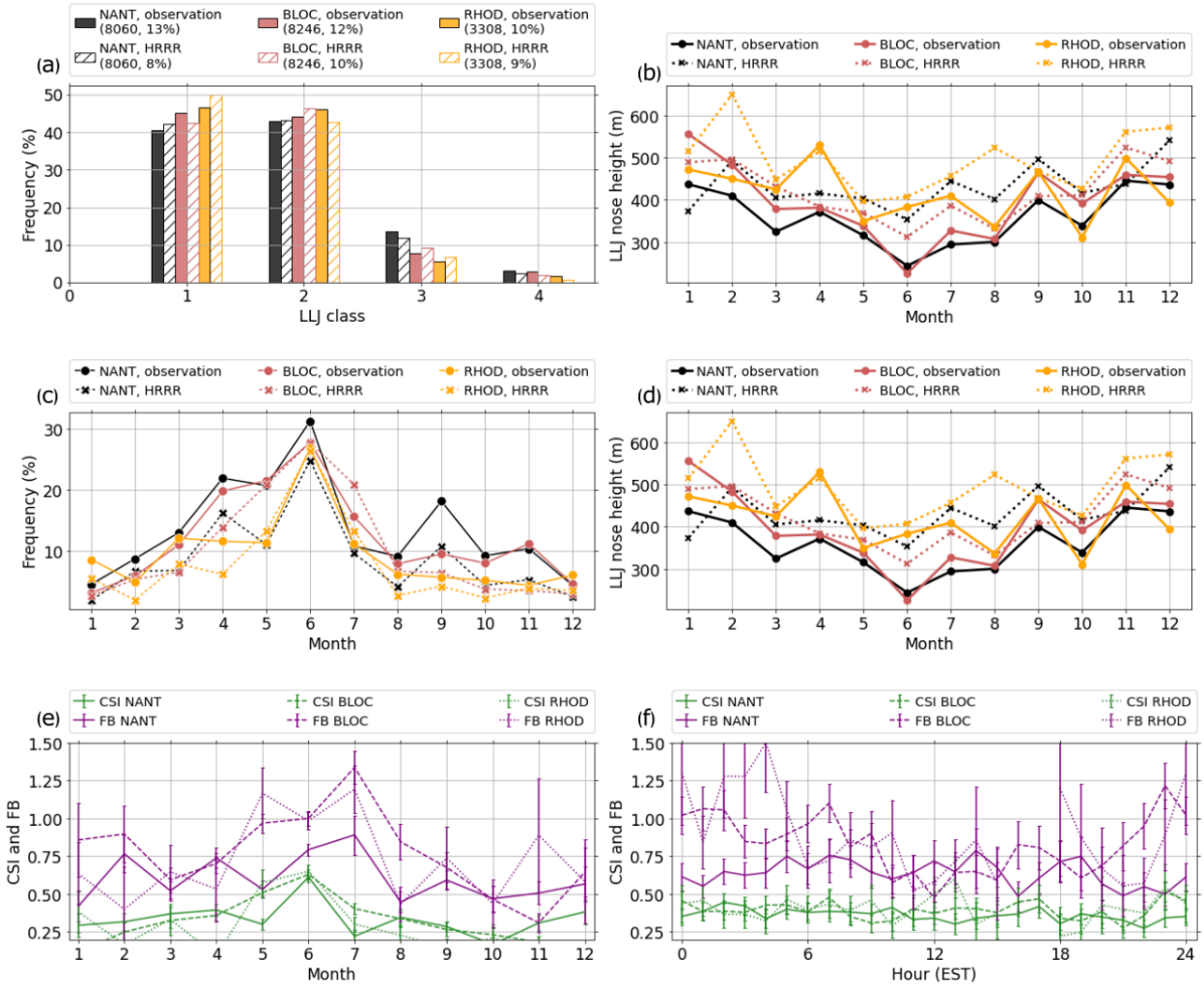

Fig. 1: (a) *Relative frequency of LLJ class, (b) mean LLJ nose height per month, (c) relative frequency of LLJ occurrence per month, and (d) relative frequency of LLJ occurrence per hour of the day. The relative frequency of LLJ class in (a) is with respect to the times when a LLJ occurred and the relative frequency of LLJ occurrence in (c) and (d) is with respect to the number of valid profiles per month (c) and hour (d) that were available for LLJ detection. CSI (Eq. 2) and FB (Eq. 3) of the HRRR model per (e) month and (f) hour of the day at NANT, BLOC, and RHOD. HRRR model data for forecast hour 12 are shown. A minimum nose wind speed threshold of 10 m/s is required (LLJ classes 1-4).*

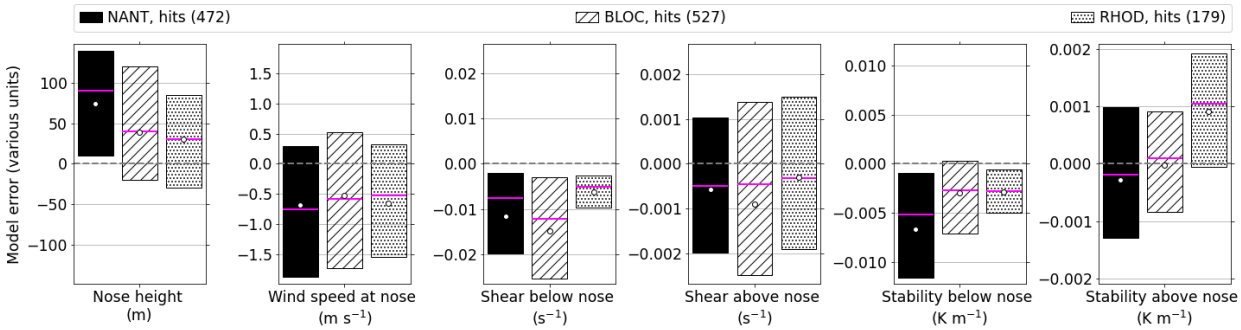

Fig. 2: Box plot of HRRR model errors (model minus observation) at forecast hour 12 of LLJ nose height, wind speed at nose height, shear below nose height, shear above nose height, static stability below nose height, and static stability above nose height (from left to right) during LLJ hits at NANT, BLOC, and RHOD. The white circles indicate the mean biases, and boxes show the interquartile range with the median indicated by the horizontal pink line. Error units are given in brackets below each subplot. A minimum nose wind speed threshold of 10 m/s is required (LLJ classes 1-4).

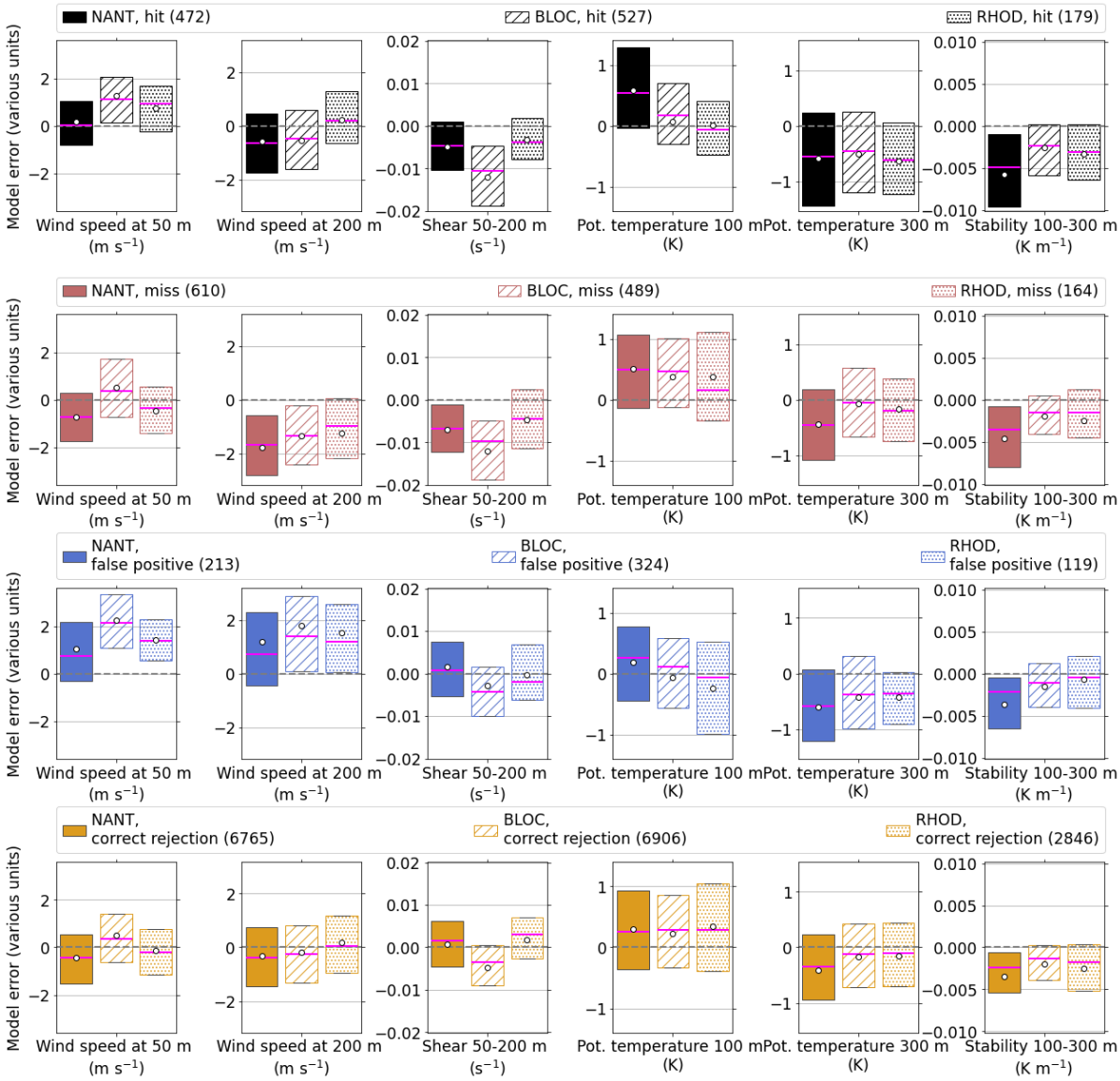

Fig. 3: Box plot of HRRR model errors (model minus observation) at forecast hour 12 of wind speed at 50 and 200 m, wind shear between 50 and 200 m, potential temperature at 2 and 300 m, and virtual potential temperature difference (stability) between 2 and 300 m (from left to right) during LLJ (a) hits, (b) misses, (c) false positives, and (d) correct rejections. The white circles indicate the mean biases, boxes show the interquartile range with the median indicated by the horizontal pink line. Error units are given in brackets below each subplot. A minimum nose wind speed threshold of 10 m/s is required (LLJ classes 1-4).

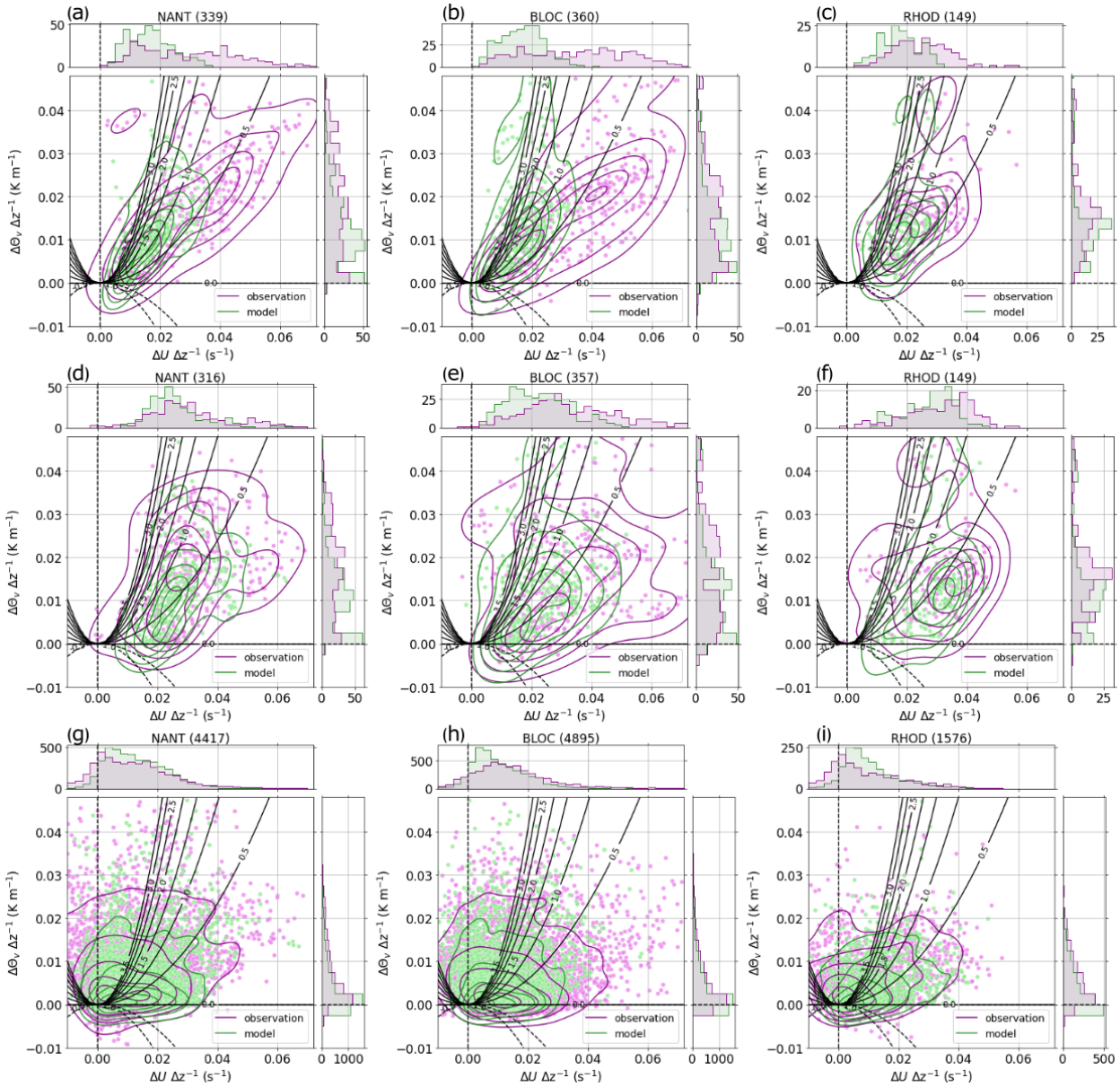

Fig. 4: (a–c) Relationship between low-level static stability and horizontal wind shear for LLJ hits for gradients computed up to the LLJ nose height in the observations (purple) and HRRR model (green) at NANT, BLOC, and RHOD. Relationship between low-level static stability (between 100 m and 300 m) and horizontal wind shear (between 50 m and 200 m) during (d–f) LLJ hits and (g–i) LLJ correct rejections. The number of samples is given in brackets above each subplot. Marginal axes show histograms. The black lines indicate theoretical Bulk Richardson numbers. HRRR data for forecast hour 12 are shown. A minimum nose wind speed threshold of 10 m/s is required (LLJ classes 1-4).

3. The analysis of bulk Richardson number is a strong part of the manuscript. The finding that HRRR underrepresents low-Ri regimes during LLJ hits is important because it links wind and thermodynamic errors to the dynamic stability of the marine boundary layer. However, the physical implications could be explained more clearly. For instance, if the model overestimates Ri because it underestimates shear more strongly than it underestimates static stability, this could imply that the model boundary layer is dynamically too stable, potentially affecting turbulent mixing, rotor-layer structure, wind-energy-relevant shear, and LLJ maintenance. The authors should briefly discuss how this error might influence forecast applications and model physics development.

We added the following discussion to the end of Section 4.3 on Bulk Richardson number regimes:
The model discrepancies in representing correct Bulk Richardson number regimes likely originate from two primary factors: (1) suboptimal regulation of the Prandtl number, and (2) insufficient calibration of near-surface mixing intensity relative to surface stability. Regarding the former, the stability functions for momentum (SM) and heat (SH) within the MYNN-EDMF framework govern the proportional mixing of momentum and heat, as expressed by the Prandtl number (Pr = SM/SH). When Ri > 0.1, SM exceeds SH and can trend toward infinity as Ri surpasses 1. Refined management of momentum mixing in this regime is expected to mitigate the overestimation of weak shear and the simultaneous underestimation of strong shear. Concerning the latter factor, the magnitude of near-surface mixing in the MYNN-EDMF is chiefly governed by the surface layer length scale (ls), which is sensitive to the surface stability parameter (ζ = z/L). Here, L represents the Obukhov length (e.g., Stull, 1988) and z denotes the height above ground level. It is plausible that the regulation of ls requires distinct treatment over marine versus terrestrial surfaces. Alternatively, the estimation of ζ could be re-evaluated using flux-profile relationships specifically developed for stable marine environments. Continued investigation into these mechanisms may address the systematic underestimation of shear and LLJ intensity, as well as the positive bias in jet maximum heights.

4. The discussion of HRRR errors in low-level temperature, static stability, and wind shear could be strengthened by briefly considering uncertainties in surface energy budget representation. The manuscript mainly attributes some near-surface errors to SST and land-sea thermal contrast, which is reasonable. However, other surface energy exchange processes may also influence boundary-layer structure, especially under precipitation conditions. Studies (Gillett & Cullen. 2011; Zhou et al. 2024a,b) have shown that precipitation-induced surface sensible heat flux, a process often neglected in weather and climate models, can modify surface energy partitioning and affect regional simulations. I suggest that the authors briefly discussing broader uncertainties in surface energy budget processes and model physics.

Gillett, S., & Cullen, N. J. (2011). Atmospheric controls on summer ablation over Brewster Glacier, New Zealand. International Journal of Climatology, 31(13), 2033–2048.
Zhou, X., Ray, P., Tan, H., Dudhia, J., Ajayamohan, R. S., Gomes, H., & Pan, Y. (2024). Rain‐induced surface sensible heat flux reduces monsoonal rainfall over India. Geophysical Research Letters, 51(14), e2023GL107796

Zhou, X., Ray, P., Dudhia, J., Tewari, M., Nikolopoulos, E., Johnson, N. C., & Hagos, S. (2024). On the importance of precipitation‐induced surface sensible heat flux for diurnal cycle of precipitation in the maritime continent. Geophysical Research Letters, 51, e2024GL111940.

We now briefly describe uncertainties related to the surface energy budget at the end of Section 4.3:
Other sources of errors could be associated with the bulk surface flux algorithm used over water (COARE3.0; Fairall et al. 2003), the detailed representation of the island soil temperature and moisture, and the representation of precipitation-induced surface sensible heat fluxes (Zhou et al. 2024a,b).

Fairall, C. W., E. F. Bradley, J. E. Hare, A. A. Grachev, and J. B. Edson, 2003: Bulk parametrization of air–sea fluxes: Updates and verification for the COARE algorithm. J. Climate, 16, 571–591.

Minor comments:
1. The abstract is clear and informative, but the statement that SST errors partly explain the weakened stability should be phrased cautiously because the evidence is site-dependent and correlations are weak to moderate.

We modified the statement to: *However, low-level horizontal wind shear and static stability were too weak in the model, especially during the warmer months. At certain sites, these biases were accompanied by errors in sea surface temperature, suggesting a potential, though localized, link.*

2. In Sect. 3.2, the authors show that wind speed standard deviation increases with forecast hour, while mean biases remain relatively stable. This is an interesting result and could be highlighted more clearly as evidence that forecast variability errors grow faster than systematic mean errors.

We added the following sentence: *While the bias was relatively constant after the first few forecast hours, standard deviation of wind speed at 50 and 200 m increased with forecast hour for all seasons and all sites, indicating that forecast variability errors grew faster than systematic mean errors, likely linked to the general degradation of model skill with forecast hour.*

3. The conclusion should more clearly separate robust findings from site-specific findings. For example, the underestimation of static stability appears robust across sites, while the source of shear error differs by site.

We rewrote parts of the summary to be more specific about the wind speed and shear errors that are site-specific: *Mean biases in low-level wind speed were generally less than 1 m/s and did not grow much for longer forecast hours. They did, however, vary in sign by site with an underestimation at NANT, the site furthest away from the coast, and an overestimation at RHOD at the coast. Low-level shear errors were often small on the average at NANT and*

RHOD especially during the colder seasons, but consistently underestimated at BLOC due to an overestimation of wind speed at 50 m, likely related to the erroneous representation of the island topography.