Comment on Ricciardulli et al.'s (2025, BAMS):

Investigation of a Calibration Change in the Ocean Surface Wind Measurements from the TAO Buoy Array

Collin A. Cruz, Mark A. Bourassa, Shawn R. Smith

Center for Ocean-Atmospheric Prediction Studies (COAPS)

Supported by:

The TAO Buoy Array & the Problem

What Is TAO?

- Tropical Atmosphere Ocean (TAO)
- 55 moored ocean buoys in the tropical Pacific (165°E–95°W, 8°S–8°N)
- Monitors ENSO, tropical climate variability
- Contributes to satellite calibration as ground comparison data

The Problem

- Starting mid-2020, satellite–buoy wind differences jumped 0.5–0.8 m/s
- Seemingly coincides with anemometer replacement
- 44 of 55 buoys serviced Mar–Sep 2020

Why Does This Matter?

- Satellite wind retrievals are calibrated against buoy winds — a buoy bias propagates everywhere
- Affects air–sea flux estimates (evaporation, momentum, heat), NWP models, climate records
- The 0.5–0.8 m/s step change nearly matches the typical tropical Pacific yearly wind anomalies (~0.8 m/s)

Ricciardulli et al. (2025) — Key Findings

~9%
Average systematic wind speed increase

0.5–0.8
m/s bias vs. GMI, ASCAT, WindSat, AMSR2

55/55
TAO buoys now affected (all serviced by 2024)

How They Found It

- Collocated TAO buoy winds with multiple satellite wind datasets (GMI, ASCAT-A/B/C, WindSat, AMSR2) within 25 km / 30 min
- Both satellite and buoy winds converted to 10m reference height
- Partitioned data pre/post each buoy's servicing date — every satellite showed the same jump
- Suggested likely cause: change in wind tunnel calibration lab for new anemometers

Our Contribution: Ship-Based Verification

Can we independently confirm the bias using a completely different platform?

Ricciardulli et al.

- Satellite vs. Buoy collocations
- Remote sensing (GMI, ASCAT, etc.)
- ~65,000 collocations per period
- Winds at 10-m reference height

This Study

- Ship vs. Buoy collocations
- In situ - Shipboard Automated Meteorological and Oceanographic System (SAMOS) research vessels)
- 831 collocations (635 pre, 170 post)
- Winds at buoy height (4.0 m)

Data & Methods

Ship Observations

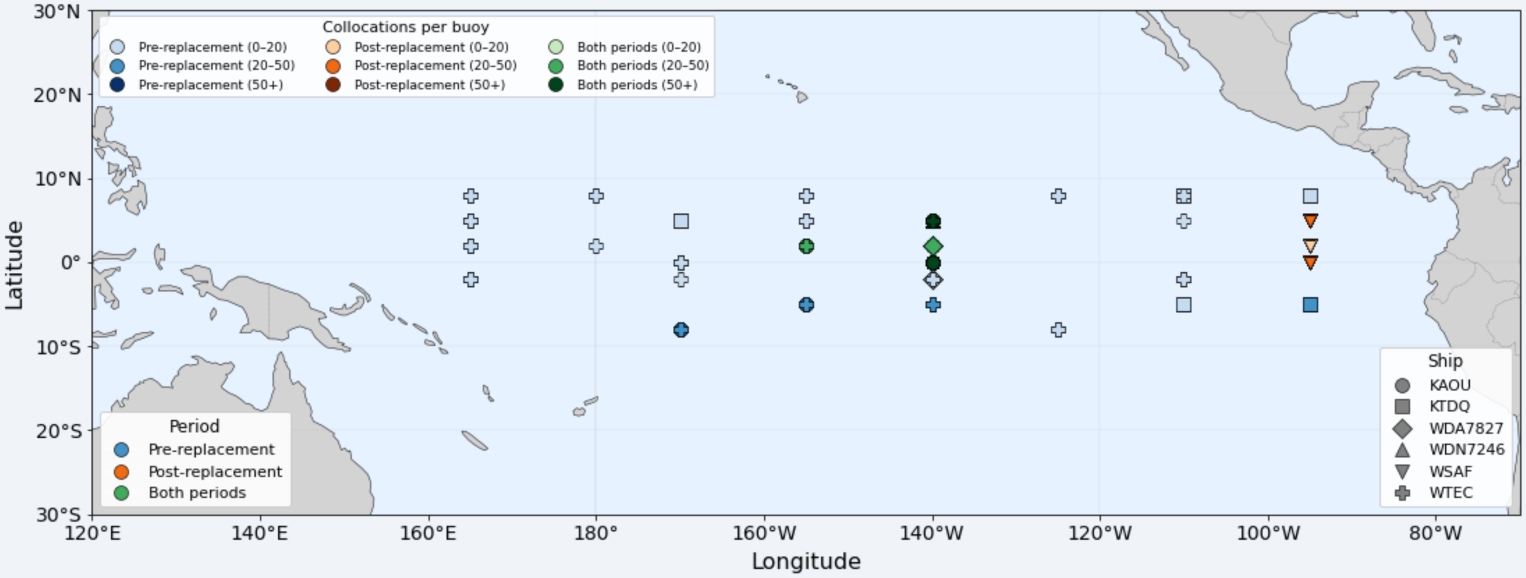
- 6 SAMOS vessels
- 1-min data → 10-min averages
- Period: 2010–2025

Quality Control

- SAMOS QC flagging, NDBC QC, height adjustment QC
- Ship winds → buoy height via Monin-Obukhov similarity
- All used buoys → 4.0 m

Collocation Criteria

- Ship within 25km of buoy
- Pre/post split per Ricciardulli et al. Table C1
- Difference = buoy wind – ship wind



Collocation locations — marker shape = ship, color = period, shade = collocations

Two-Tier Height Adjustment QC System

Critical Flags → EXCLUDED

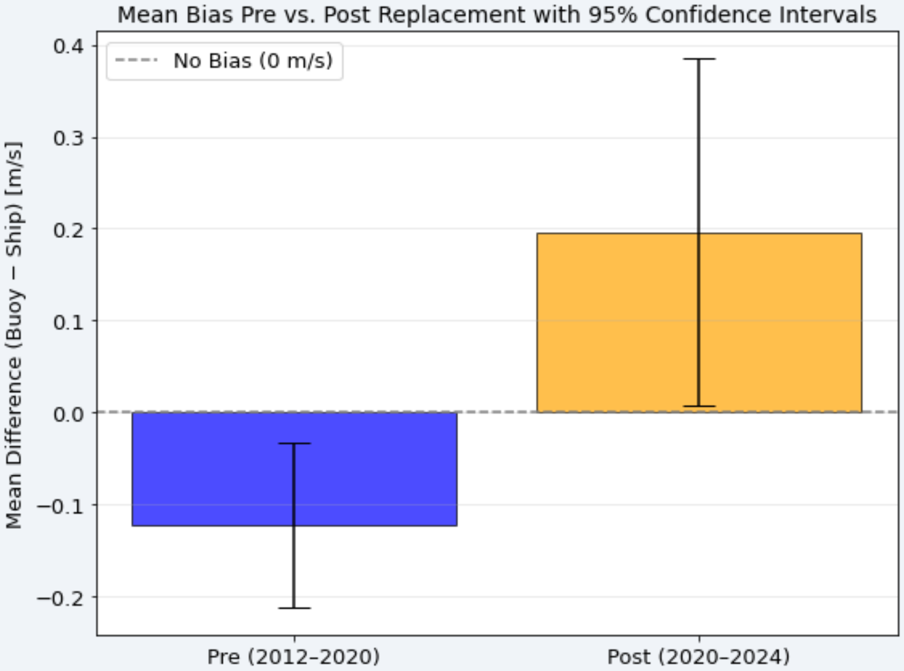
- Missing or out-of-range wind speed
- Invalid ship anemometer height
- Missing air temperature or SST
- *These parameters strongly affect the height adjustment*

Soft Flags → RETAINED

- Missing RH → default 80%
- Missing SLP → default 101,325 Pa
- Value clamped to valid range
- *Minor influence on stability correction*

Additional: 13 records excluded where buoy read 0 m/s but ship read >2.0 m/s (likely instrument error)

Result: Pre- vs. Post-Replacement Bias



PRE (n = 661)

-0.122 m/s

± 0.046 m/s | p = 0.008 | Ships slightly higher

POST (n = 170)

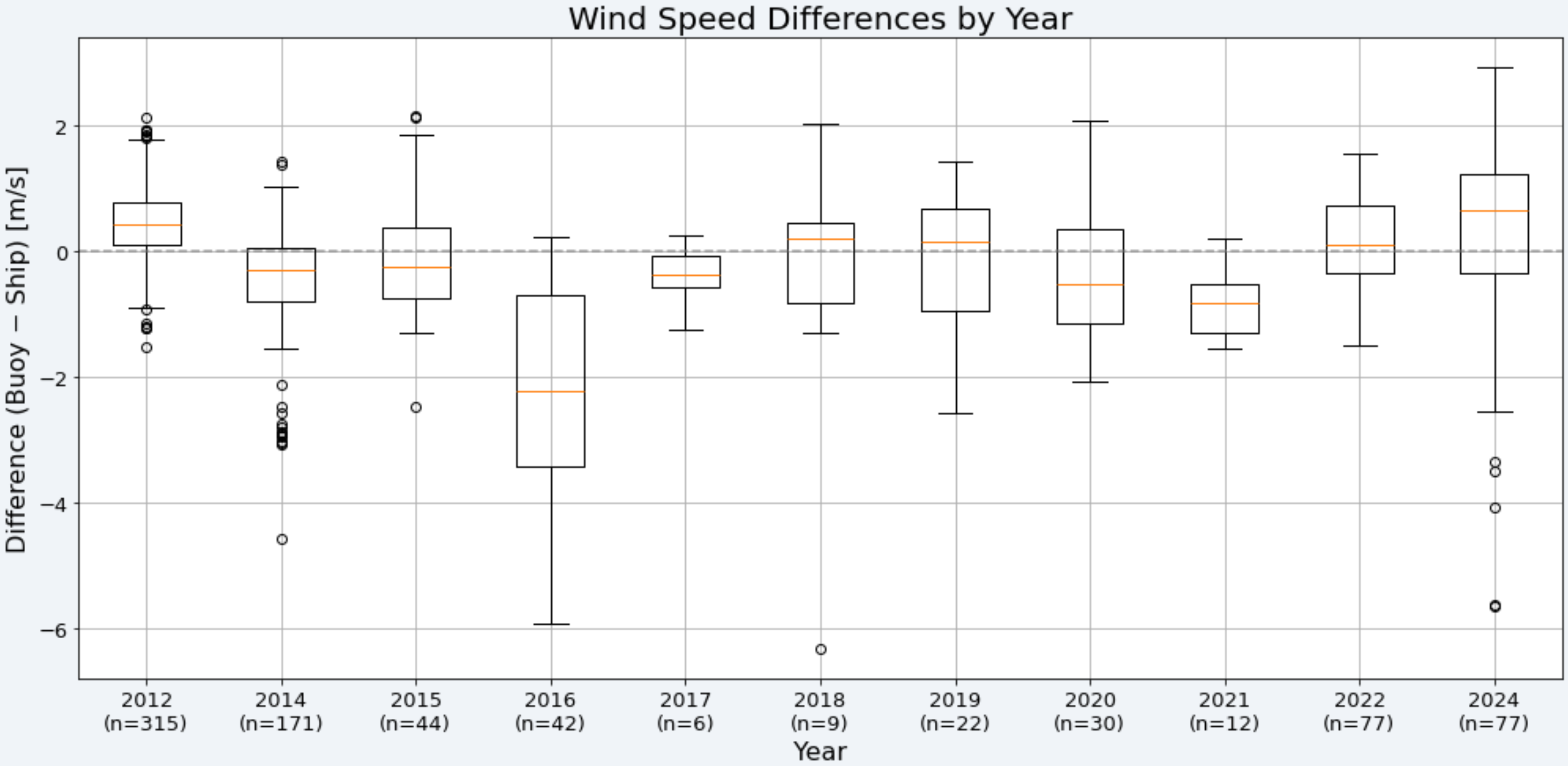
+0.196 m/s

± 0.095 m/s | p = 0.041 | Buoys now reading higher

Total Shift: ~0.32 m/s

- Smaller than Ricciardulli et al.'s 0.5–0.8 m/s, but identical in sign and timing
- Both means statistically significant (one-sample t-test, $\alpha = 0.05$)

Result: Interannual Variability



Large year-to-year variability with outlier years: 2016 (median +2.23 m/s) and 2024 (median −0.66 m/s)

Most years cluster near zero — reinforces need for multi-year (≥3–5 yr) averaging as ground truth

Statistical Significance by Year

Year	n	Median (m/s)	Corrected p	Sig?
2012	315	**0.434**	0.002	*
2014	171	-0.285	0.9722	
2015	44	-0.235	1.000	
2016	42	**-2.229**	0.000	*
2020	30	-0.514	1.000	
2021	12	-0.816	0.166	
2022	77	0.095	0.936	
2024	77	**0.658**	0.046	*

Reference: 2017–2019 (n = 37, median 0.010 m/s)

METHOD

Each year vs. baseline

Reference: 2017–2019

Mann-Whitney U test

Bonferroni-corrected, α = 0.05

3 / 8

tested years differ from baseline

OUTLIER YEARS

2012 −0.43 m/s

2016 +2.23 m/s

2024 −0.66 m/s

Outliers are isolated — the 2017–2019 baseline holds, supporting multi-year averaging as ground truth.

Why Is Our Magnitude Smaller?

Sign and timing agree; magnitude differs (~0.32 vs 0.5–0.8 m/s). Likely factors:

1. **Reference Frame: Current-Relative vs. Earth-Relative Winds**
   Satellites read winds relative to the moving sea surface; ships and buoys read earth-relative. Equatorial currents (~0.2–0.5 m/s) inflate satellite–buoy differences.

2. **Reference Height: 4 m vs. 10 m and Bias Amplification**
   Ricciardulli et al. adjust to 10 m, assuming a speed-dependent bias, that bias may be amplified by log-profile extrapolation; we stay at 4 m, leaving it unscaled.

3. **Sampling: Coverage and Sample Size**
   Satellites: ~65,000 collocations per period, all 55 buoys. Ships: 831 (170 post) from 6 SAMOS vessels, 35 buoys.

4. **Platform Effects: Ship Flow Distortion and Residual Bias**
   Ship anemometers carry residual superstructure flow distortion that satellites do not. Per-vessel offsets can partly mask the buoy signal.

Discussion & Takeaways

1

Independent Confirmation

Ship-buoy collocations corroborate the satellite-detected bias. The sign and timing match Ricciardulli et al.

2

Interannual Variability

Large year-to-year scatter (especially 2016, 2024) reinforces the need for multi-year averaging (≥3–5 years) when using TAO data as ground truth.

3

TAO Remains Valuable

The array is still a critical data source — provided users account for the post-replacement offset and use sufficiently long averaging periods.

Questions?

Acknowledgements:

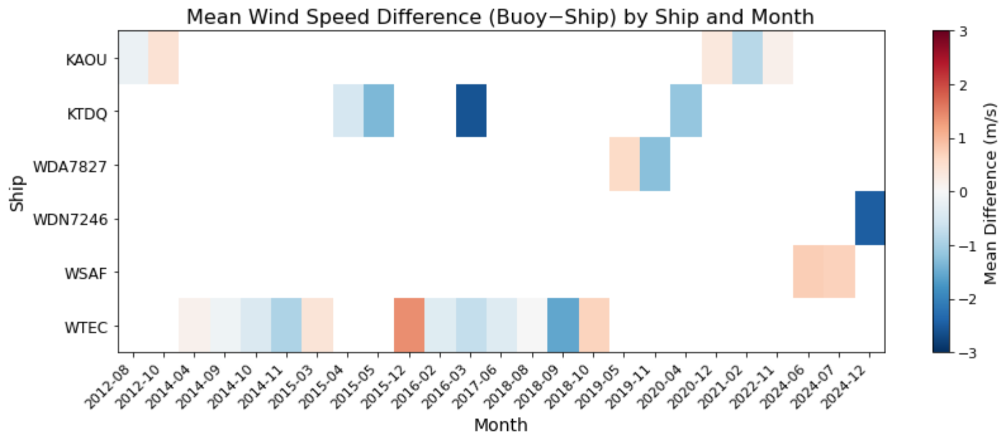
This analysis was supported by NSF Oceanographic Instrumentation and Technical Services Program, grant OCE-2439640.
MAB's support is from NASA PO for OVWST leadership.
I would also like to acknowledge the GTMBA Project Office of NOAA/PMEL for their data/QC.

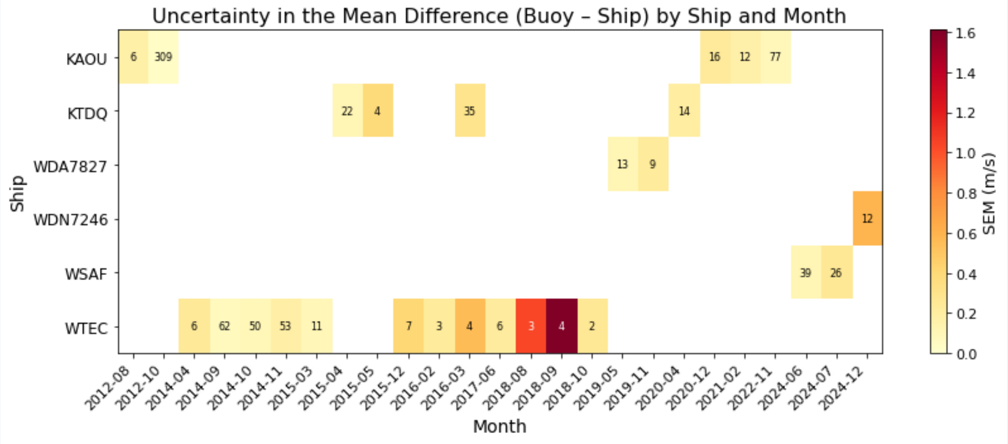
Appendix

Backup Slides for Q&A

Methods, robustness checks, and follow-up answers

Result: Ship-Month Heatmaps





No systematic ship-specific bias.

- Differences are distributed across all ships and time periods.

- Darker cells indicate higher uncertainty, driven primarily by fewer collocations in that month.
- Pre-replacement months (especially WTEC 2016–2018) show elevated SEM, while post-replacement cells are generally lighter and more uniform, reflecting more consistent data.

Methods Note: Statistical Tests

How the per-year significance results were computed

THE TEST

Mann–Whitney U

What it does

Tests whether two samples come from the same distribution by comparing ranks, not means.

Why use it here

Non-parametric — no normality assumption. Robust to outlier years like 2016.

How applied

Each year's differences vs. the 2017–2019 baseline.

THE CORRECTION FOR FALSE POSITIVES

Bonferroni Correction

The problem

Run 8 tests at $\alpha = 0.05$ and ~40% chance you flag a year by chance alone.

The fix

Multiply each p-value by the number of tests (8). Capped at 1.000.

The trade-off

Conservative — may miss small real effects.

If a year still shows $p < 0.05$ after this, the effect is real.

Backup: Sample Size & Statistical Power

Why 170 post-replacement collocations are sufficient to detect the bias

THE SETUP

661 / 170

Pre / post collocations.
6 ships × 35 buoys × 13 years.

THE TEST

$p = 0.041$

One-sample t-test on post.
Mean ≠ 0 even with n = 170.

Why a small post-sample is still informative

- **Independence is the point — not n.** Ship and satellite errors are uncorrelated; the matching sign + timing across two unrelated platforms is the diagnostic, not raw count.
- **Why post is small.** Pandemic-era reduction in research-vessel transits, plus a shorter post-window (~4 yr vs. ~8 yr pre).
- **What we don't claim:** a precise magnitude. The 95% CI on the post-mean is wide; n = 170 detects the bias but doesn't pin it to ±0.05.

Backup: Currents & Reference Frame

How much of the magnitude gap is currents alone? A back-of-envelope estimate

THE PHYSICS

Satellites measure wind relative to the moving sea surface

Satellite	Buoy / Ship	Difference
$U_{rel} = U_{air} - U_{cur}$	$U_{abs} = U_{air}$	$\Delta U \approx U_{current}$
Sees relative wind	*Sees earth-relative wind*	*0.2–0.5 m/s near 0°*

WHAT THE NUMBERS LOOK LIKE

Equatorial South Equatorial Current (near 0°): westward, 0.3–0.5 m/s typical.

Off-equator (5–8° N/S): currents drop to 0.1–0.2 m/s, biases shrink toward true wind difference.

Implied magnitude correction

If satellites overestimate the buoy step by ~0.2 m/s due to current-relative framing, **Ricciardulli's 0.5–0.8 m/s collapses toward our 0.3 m/s.** The two studies are quantitatively consistent.

Backup: Why We Stay at 4 m (Not 10 m)

Reference height affects bias magnitude — a key reason our number is smaller

RICCIARDULLI ET AL.

Adjust both to 10 m

Buoy: 4 m → 10 m

Satellite: equivalent-neutral 10 m wind

Log-profile extrapolation amplifies any speed-dependent bias by *ln(10/4) / ln(z_0-implied) ≈ 1.10–1.15 for tropical winds.*

→ a 0.5 m/s 'bias' at 10 m can correspond to ~0.43 m/s at 4 m.

THIS STUDY

Match buoy height (4 m)

Ship anemometer → 4 m

Buoy: native 4m

Bringing both to a common *low* height keeps the bias at its raw, unscaled magnitude.

Bottom line

Magnitudes between the two studies are not directly comparable — the height choice already explains a fraction of the gap before we get to currents, sampling, or platform effects.

Backup: Why Pre-Replacement Isn't Zero

Interpreting the −0.122 m/s pre-replacement bias

PRE-REPLACEMENT

−0.122 m/s

Buoy minus Ship.
Ships read slightly higher.

THREE PLAUSIBLE CONTRIBUTIONS

- **Residual flow distortion.**
  Even SAMOS-quality vessels carry hull/superstructure effects that bias true winds slightly high. Magnitude is platform-specific; ~0.1–0.2 m/s is typical.
- **Height-adjustment uncertainty.**
  Ships at 15 m brought to 4 m via Bourassa MFT carry a small residual error from imperfect Tair / SST / RH. Bias direction depends on stability.
- **Buoy under-read at low winds.**
  The pre-replacement TAO anemometers are known to slightly under-respond near calm conditions, biasing the buoy low. Pre period had far more lower wind speed observations than post.

Key point: the small pre-bias is the baseline disagreement between platforms — what matters is the change from −0.12 to +0.20 m/s.

Backup: Anemometer Changes — Confounding?

Per-vessel calibration history vs. the buoy step change

THE CONCERN

Confounding?

If a ship swapped anemometers around 2020, that swap — not the buoy — could create the apparent step change.

THE EVIDENCE

Six ships, many dates

WTEC: 2014, 2019, 2020-11, 2021, 2025
KAOU: 2013, 2020-10, 2022, 2024
WSAF: 2020-05, 2021, 2022 (×2), 2023, 2024
WDA7827: 2022, 2025
WDN7246: 2024
KTDQ: 2013, 2021, 2023, 2024 (×2), 2024

THE CONCLUSION

Not the cause

Six ships, asynchronous swap dates, no shared event — yet the pre→post step appears in the aggregate.

Random per-ship offsets average toward zero; only a common signal (the buoys) survives.

Heatmap evidence

The ship-month bias heatmap (slide 13) shows no single ship dominates the post-replacement signal — biases shift sign across vessels and months, ruling out a per-ship instrument swap as the driver.

Backup: The 2016 Anomaly

Why we retain 2016 in the analysis: QC discipline, not post-hoc filtering

WHAT 2016 LOOKS LIKE

Total collocations	42
Mean difference	−2.26 m/s
Dominant ship	KTDQ (83%)
Dominant buoy	5s95w (83%)
Concentrated date	2016-03-15

WHY WE DON'T EXCLUDE IT

- **QC was set prior.**
  Same SAMOS letter flags + range rules applied to every year.
- **No new rule for 2016.**
  Excluding it requires a physical rule, not a result-driven one.
- **Outliers ≠ wrong.**
  QC should reject the implausible, not the merely unusual.
- **Mann-Whitney handles it.**
  Rank-based tests already down-weight tail-heavy years.

Transparent reporting: the result with and without 2016

- **With 2016 (reported):** pre = −0.122, post = +0.196, step = 0.3 m/s.
- **Without 2016 (sensitivity):** pre = +0.023, post = +0.196, step ≈ 0.2 m/s. *Step is smaller but stays positive — direction agrees with Ricciardulli either way.*

Backup: Why Use the Bourassa MFT?

Stability-aware over neutral log-profile — and what it costs

THE SIMPLE OPTION

Neutral log profile

$$u(z) = u_0 \cdot \ln(z / z_0) / \ln(z\_ref / z_0)$$

Pros: fast, no thermo data needed.
Cons: assumes neutral stability. The tropical Pacific ABL is rarely neutral — afternoon convection is the rule.

Error in unstable conditions can reach 0.2–0.3 m/s for 15 m → 4 m extrapolation.

WHAT WE USE

Bourassa MFT (v23v8)

Iterative Monin-Obukhov similarity with sea-state-dependent z_0.

Inputs: U, Tair, SST, RH, SLP, net LW.
Stability: computed online, not assumed.

Why it matters: the bias signal we hunt (~0.3 m/s) is the same magnitude as neutral-profile error in unstable conditions.

Validation in our pipeline

MFT is called row-by-row; on rare exceptions the code falls back to neutral log profile and the row is flagged **critical** (excluded by default). Less than ~0.1% of rows trigger this fallback.