

S1 Additional methods and validation results

S1.1 DJI log decryption and data extraction

All flights used aircraft firmware v01.00.1500. Telemetry from each flight was extracted from the proprietary DJI FlightRecord files using DJI log parser v0.5.7 (Vauvillier, 2025). The procedure and the fields extracted from the original flight logs are listed below.

Extraction procedure:

1. Obtain dji-log binary (included at tools/decryption/dji-log-parser/dji-log).
2. Set the application programming interface (API) key: `export DJI_API_KEY="<key>"`.
3. Decrypt: `dji-log <file>.txt --csv <file>.csv --api-key "$DJI_API_KEY"`.

The decryption step requires an API key for the third-party dji-log service and is therefore not fully self-contained. All processed products used in the analysis are included in the archive, so reproduction can begin from those records when an API key is unavailable.

Extracted fields:

- `CUSTOM.dateTime`: Recorded timestamp with millisecond precision.
- `OSD.pitch`, `OSD.roll`, `OSD.yaw`: Attitude angles (degrees).
- `OSD.latitude`, `OSD.longitude`: World Geodetic System 1984 (WGS84) horizontal coordinates.
- `OSD.altitude`: Altitude recorded in the flight log (m).
- `RC.rudder`: Yaw stick deflection (-1024 to +1024).

Timestamp alignment: The recorded drone timestamps (`CUSTOM.dateTime`) and TriSonica logger timestamps (Polle, 2026) were converted to Coordinated Universal Time (UTC), and the two series were merged by linear interpolation onto a common 10 Hz grid without an additional time-offset correction. The retained aligned files (`data/processed/aligned_10hz/*_aligned.csv` in the archive) are reduced analysis products. They contain the timestamp, pitch, roll, yaw, and rudder-control fields, together with reference wind speed and direction (`wind_speed` and `wind_direction`). The derived body-frame horizontal components are `wind_u_body` and `wind_v_body`. These files do not contain aircraft altitude, vertical wind, sonic temperature, or TriSonica diagnostic flags. No separate fuller TriSonica export is included in the archive. Effects of vertical motion therefore cannot be evaluated from these products (Sect. 4.2 of the main article). Yaw and wind direction were unwrapped before interpolation to prevent 0/360° discontinuities and then wrapped back to their original range. Missing samples were handled by the window-level completeness criterion (Table S1).

A timing audit used cross-correlation between the held-out Ridge and reference wind-speed traces. The diagnostic therefore depends on the prediction trace and is not an independent clock measurement. Lags at peak correlation ranged from -1 to +4 s in 49 sessions and had a median magnitude of 1 s. The remaining session had a lag magnitude of 38 s (Sect. S1.14). The audit was performed after the analysis and no lag correction was applied. It is reproduced by `scripts/timing_audit.py`, which writes `data/results/timing_audit.csv`.

S1.2 Reference sensor alignment procedure

The TriSonica Mini was aligned to true north using the mobile compass application by PixelProse SARL. The application was operated in true north mode. The DJI yaw telemetry provided the aircraft heading relative to north. Any remaining constant offset between the two is included in the 5–7° uncertainty range for the reference direction.

Equipment:

- Digital magnetometer (Android compass application by PixelProse SARL).
- Tripod with bubble level and 90 cm mast extension.
- Physical directional marker aligned to sensor body frame.

Alignment procedure:

1. Level tripod ($\pm 0.5^\circ$).
2. Mount sensor with forward axis marked.
3. Align tripod assembly until magnetometer reads 0° to true north.
4. Verify repeatability after disturbances.

Uncertainty budget: The combined root-sum-square (RSS) uncertainty of the reference direction accounts for:

- Magnetometer accuracy: $\pm 2\text{--}3^\circ$.
- Mounting precision: $\pm 1^\circ$.
- Magnetic declination variability: $\pm 0.5^\circ$.
- Sensor housing asymmetry: $\pm 1^\circ$.
- Instrument direction accuracy: $\pm 1.0^\circ$ (manufacturer specification).

Combining these as a root sum of squares with the 3° magnetometer limit gives $\sqrt{3^2 + 1^2 + 0.5^2 + 1^2 + 1^2} = 3.5^\circ$ for the reference-sensor alignment and instrument term. Combining that term with the aircraft heading uncertainty of $1\text{--}2^\circ$ gives a quantified root-sum-square range of $3.6\text{--}4.0^\circ$. The wider $5\text{--}7^\circ$ range is a conservative allowance for unquantified mounting and alignment effects, local flow distortion, and the finite separation between the aircraft and reference sensor. It is not a formally propagated uncertainty.

S1.3 Flight site coordinates

The dataset was collected at two locations in the Bremen area:

- **Ground-level site:** $53^\circ 06' 08.9''\text{N}$ $8^\circ 50' 59.0''\text{E}$ (open field).
- **Hill site:** $53^\circ 07' 09.7''\text{N}$ $8^\circ 48' 26.1''\text{E}$ (elevated terrain).

S1.4 Quality control filter specifications

Quality-control (QC) filters were designed to exclude periods that do not represent stationary hover rather than to select a wind regime. No meteorological wind-speed threshold was applied, and neither altitude nor horizontal position was an explicit filter.

Table S1 lists the filters and thresholds used for the analysis.

After filtering, the analysis dataset contains 71,749 rows and 62 features. Each row represents one rolling 30 s window. Within continuous retained segments, successive rows advance by 1 s and share 29 s of raw samples. Quality-control gaps and session boundaries interrupt this sequence. The row count therefore corresponds to 19.9 h of window-equivalent time at the analysis step, while the union of retained window intervals spans 22.5 h. The overlapping windows are not statistically independent, which is why whole sessions form the unit of cross-validation.

Table S1. Quality-control filters. No meteorological wind-speed threshold was applied. Session-mean winds span 0.8–7.6 m s⁻¹. Altitude and horizontal position were not filtered (Sect. 2.3 of the main article).

Filter	Specification
Trim	Mask the first and last 10 s of each session.
Yaw exclusion	Exclude intentional manoeuvres, defined by a yaw-stick deflection $ RC.rudder \geq 40$ counts after values with $ RC.rudder \geq 900$ are treated as unavailable. Stick-active segments are padded by ± 2 s and extended by a 30 s-wide exclusion window (± 15 s). Retained windows may not overlap any excluded sample.
Continuous duration	Remove retained intervals shorter than 30 s after trimming and yaw exclusion.
Completeness	Interpolate the drone and reference series onto the 10 Hz grid, then drop 30 s windows in which $< 80\%$ of samples have finite values.
Reference validity	Drop any 30 s window containing a negative TriSonica speed value used to mark invalid data.

S1.5 Complete feature list

All 62 model features are derived from pitch and roll. Some summarize the samples within each 30 s window, while others describe changes across the retained sequence of windows. Table S2 groups them by their statistical or temporal role and provides the name used in the released feature matrix. Yaw is excluded because it is used only for the body-frame transformation.

Table S2. Complete list of 62 features derived from pitch and roll using 30 s rolling windows with a 1 s step. Yaw is used only for coordinate transformation and is excluded from model inputs.

Category	Feature name	Description
Basic statistics		
	pitch_mean	Mean pitch (degrees)
	pitch_std	Std dev of pitch (degrees)
	pitch_min	Minimum pitch (degrees)
	pitch_max	Maximum pitch (degrees)
	roll_mean	Mean roll (degrees)
	roll_std	Std dev of roll (degrees)
	roll_min	Minimum roll (degrees)
	roll_max	Maximum roll (degrees)
Simple nonlinear terms		
	pitch_roll_mean	$\langle \theta \phi \rangle$, mean product of paired pitch and roll samples
	abs_pitch_mean	Mean $ \text{pitch} $
	abs_roll_mean	Mean $ \text{roll} $
	pitch_squared_mean	$\langle \theta^2 \rangle$, mean squared pitch
	roll_squared_mean	$\langle \phi^2 \rangle$, mean squared roll
Temporal lags		
	pitch_lag_1s	Pitch mean one retained window earlier
	roll_lag_1s	Roll mean one retained window earlier
	pitch_lag_2s	Pitch mean two retained windows earlier
	roll_lag_2s	Roll mean two retained windows earlier
	pitch_lag_5s	Pitch mean five retained windows earlier
	roll_lag_5s	Roll mean five retained windows earlier
Response dynamics		
	pitch_rate	Difference of consecutive retained pitch means divided by their time interval (deg s^{-1})
	roll_rate	Difference of consecutive retained roll means divided by their time interval (deg s^{-1})
	pitch_accel	Difference of consecutive pitch rates divided by the current time interval (deg s^{-2})
	roll_accel	Difference of consecutive roll rates divided by the current time interval (deg s^{-2})
	pitch_rate_std	Trailing three-value standard deviation of pitch rate (deg s^{-1})
	roll_rate_std	Trailing three-value standard deviation of roll rate (deg s^{-1})
Coupling/geometry		
	tilt_magnitude	$\sqrt{\bar{\theta}^2 + \bar{\phi}^2}$ from window-mean pitch and roll
	tilt_angle	$\text{atan2}(\bar{\phi}, \bar{\theta})$ in degrees
	pitch2_roll	$\langle \theta^2 \rangle \bar{\phi}$
	pitch_roll2	$\bar{\theta} \langle \phi^2 \rangle$
	pitch2_roll2	$\langle \theta^2 \rangle \langle \phi^2 \rangle$
	pitch_std_roll_std	Pitch std $\times$ roll std
	pitch_range	Max pitch — min pitch (deg)
	roll_range	Max roll — min roll (deg)

Table S2. Continued.

Category	Feature name	Description
Asymmetry and normalization	pitch_asymmetry	$(\max + \min)/2 - \text{mean pitch}$
	roll_asymmetry	$(\max + \min)/2 - \text{mean roll}$
	pitch_cv	Pitch std / ($ \text{pitch_mean} + \epsilon$)
	roll_cv	Roll std / ($ \text{roll_mean} + \epsilon$)
	pitch_roll_sync	$\bar{\theta}\bar{\phi}/(\text{tilt_magnitude} + \epsilon)$
Exponential smoothing (EWMA)	pitch_ewma_fast	Fast EWMA of pitch mean (span = 3 windows)
	roll_ewma_fast	Fast EWMA of roll mean (span = 3 windows)
	pitch_ewma_slow	Slow EWMA of pitch mean (span = 10 windows)
	roll_ewma_slow	Slow EWMA of roll mean (span = 10 windows)
	pitch_ewma_diff	Fast — slow EWMA
Changes and local variability	roll_ewma_diff	Fast — slow EWMA
	pitch_momentum_short	Pitch mean — two-window lag
	roll_momentum_short	Roll mean — two-window lag
	pitch_momentum_long	Pitch mean — five-window lag
	roll_momentum_long	Roll mean — five-window lag
	tilt_momentum_magnitude	$\sqrt{\text{pitch momentum}_{\text{long}}^2 + \text{roll momentum}_{\text{long}}^2}$
	pitch_local_volatility	Std dev of pitch mean over the trailing 5 windows
	roll_local_volatility	Std dev of roll mean over the trailing 5 windows
	pitch_volatility_ratio	Ratio of trailing five-window to within-window pitch standard deviation
Interaction terms	roll_volatility_ratio	Ratio of trailing five-window to within-window roll standard deviation
	tilt_magnitude_squared	Squared tilt magnitude
	pitch_roll_rate_product	Pitch rate $\times$ roll rate
	pitch_roll_accel_product	Pitch acceleration $\times$ roll acceleration
	pitch_angle_rate_product	Mean pitch $\times$ pitch rate
	roll_angle_rate_product	Mean roll $\times$ roll rate
	pitch_sign	Sign of mean pitch
	roll_sign	Sign of mean roll
	pitch_abs_signed	Mean-square pitch signed by mean tilt, $\text{sign}(\bar{\theta})\langle\theta^2\rangle$
	roll_abs_signed	Mean-square roll signed by mean tilt, $\text{sign}(\bar{\phi})\langle\phi^2\rangle$

75 The signed mean-square terms $\text{sign}(\bar{\theta})\langle\theta^2\rangle$ and $\text{sign}(\bar{\phi})\langle\phi^2\rangle$ retain tilt polarity while encoding squared magnitude. The EWMA span parameters are 3 and 10 retained windows. Within continuous sequences, successive windows advance by 1 s, but the exponential weighting has no fixed cutoff in time. The averages use the recursive formulation. The volatility ratio is the trailing five-window standard deviation divided by the within-window standard deviation. In the implementation, 10^{-3} is used for coefficient-of-variation denominators and 10^{-6} for the synchronization and volatility-ratio terms. Coefficient-of-variation
80 features are capped at 10 and volatility ratios at 5.

Temporal features are calculated along the sequence that remains after QC. The matrix contains 271 gaps between successive retained windows caused by manoeuvres and invalid reference data. Resetting these features after every gap removes 1,671 subsequent windows and changes the Ridge speed RMSE from 0.3191 to 0.3196 m s^{-1} and the direction MAE from 7.013 to 7.019°. The rounded results are unchanged. This sensitivity test is reproduced by `scripts/temporal_gap_sensitivity.py`. All temporal features use only the current and earlier windows.
85

S1.6 Window-length sensitivity

The 62 features and the Ridge model were recomputed with 10, 30, and 60 s windows to test the chosen averaging interval. All three evaluations used the same 62,297 window-end times across the 50 sessions and the same LOFO folds. Restricting the comparison to common endpoints prevents changes in retained timestamps from being mistaken for a window-length effect.
90 Errors decrease as the averaging interval increases (Table S3). The 60 s window gives the lowest RMSE and MAE, while the 30 s window retains more temporal resolution than the 60 s average and remains the interval used for the main analysis. The 30 s values in this comparison differ slightly from the main results because only endpoints available at all three lengths are retained. The sensitivity test supports 30 s as a compromise between temporal resolution and error rather than as the error-minimizing choice.

Table S3. Ridge LOFO sensitivity to window length at 62,297 common window-end times. Speed RMSE and direction MAE are the mean $\pm$ standard deviation across sessions. The final column is the mean over the 39 sessions with session-mean reference wind above 2 m s^{-1} .

Window (s)	Speed RMSE (m s^{-1})	Direction MAE ($^{\circ}$)	Direction MAE, $> 2 \text{ m s}^{-1}$ ($^{\circ}$)
10	0.42 ± 0.17	8.1 ± 5.2	6.8
30	0.32 ± 0.14	7.1 ± 4.9	6.1
60	0.27 ± 0.10	6.6 ± 4.8	5.8

95 S1.7 Metric definitions and equal session weighting

The metrics operate on the headwind and crosswind components u_{ψ} and v_{ψ} and on the drone-relative wind direction D_{ψ} defined in Sect. 2.2 of the main article, with hats denoting predicted quantities. The wind speed of a 30 s window is the magnitude of its mean horizontal wind vector, $S = \sqrt{u_{\psi}^2 + v_{\psi}^2}$ for the reference and $\hat{S} = \sqrt{\hat{u}_{\psi}^2 + \hat{v}_{\psi}^2}$ for the prediction.

S1.7.1 Speed RMSE

100 The primary speed metric is the root-mean-square error of the wind-speed magnitude,

$$\text{RMSE}_S = \sqrt{\langle (\hat{S} - S)^2 \rangle}. \quad (\text{S1})$$

Angle brackets denote the average over the 30 s windows of a session. The speed RMSE measures the error in wind-speed magnitude alone, whereas the vector RMSE below combines the two component errors and is therefore at least as large.

S1.7.2 Vector RMSE

105 The vector RMSE combines the two component errors,

$$\text{RMSE}_V = \sqrt{\langle (\hat{u}_\psi - u_\psi)^2 + (\hat{v}_\psi - v_\psi)^2 \rangle}. \quad (\text{S2})$$

Unlike the speed RMSE, it remains nonzero when the predicted and reference vectors share the same magnitude but point in different directions. The two are therefore reported together: the vector RMSE carries the directional part of the error that the speed metric discards.

110 S1.7.3 Direction MAE

Direction error is wrapped circularly,

$$\text{MAE}_{\text{dir}} = \left\langle \min \left(|\hat{D}_\psi - D_\psi|, 360^\circ - |\hat{D}_\psi - D_\psi| \right) \right\rangle, \quad (\text{S3})$$

and the “ $>2 \text{ m s}^{-1}$ ” variant restricts the average to sessions whose mean reference wind exceeds this threshold. The window-level counterpart used in Sect. 3.3 of the main article instead keeps individual windows above 2 m s^{-1} and combines them
115 across all sessions.

S1.7.4 Equal session weighting

To keep longer sessions from having more influence, the LOFO and calibration-baseline metrics give each session equal weight:

1. Compute per-window error.
2. Group by hover session and compute per-session summary (RMSE for speed, MAE for direction).
- 120 3. Aggregate as the mean of the 50 session values.

S1.8 Ridge model specification

Ridge was fitted with scikit-learn version 1.7.2 (Pedregosa et al., 2011). The L_2 regularization strength λ was set to 1.0, corresponding to `alpha=1.0` in the scikit-learn interface. The 62 feature definitions were developed from physical reasoning and examination of the first 34 campaign sessions, collected from 1 to 10 October 2025. The feature set and regularization
125 strength were then kept unchanged in every validation fold. These 34 sessions remain in the analysis dataset, while the other 16 sessions were collected later and did not inform feature design.

Separate models predict u_ψ and v_ψ , and the features are standardized with statistics calculated only from the training data in each fold. Repeating LOFO with unregularized least squares changes the speed RMSE by 0.0002 m s^{-1} and the direction MAE by 0.01° . The penalty therefore has little effect on the reported errors, although it stabilizes the coefficients of correlated
130 features.

S1.9 Minimal-feature sensitivity

The comparison in Sect. 3.5 of the main article uses the same 62 features for all three machine-learning models, but it does not show how much of their difference from the calibration baseline comes from the larger feature set. To test the contribution of the feature set, a minimal Ridge model was fitted with the same LOFO folds and six primary tilt features, namely mean pitch, mean roll, and their squared and signed-squared terms. This model yields a speed RMSE of $0.33 \pm 0.13 \text{ m s}^{-1}$ and a direction
135 MAE of $7.4 \pm 7.3^\circ$. The corresponding values are $0.32 \pm 0.13 \text{ m s}^{-1}$ and $7.0 \pm 4.9^\circ$ for the full 62-feature Ridge model, and $0.39 \pm 0.15 \text{ m s}^{-1}$ and $8.4 \pm 5.0^\circ$ for the calibration baseline. The additional features mainly reduce the direction error and its between-session spread. The six-feature model retains most of the full Ridge advantage over the baseline. This comparison tests the reduced six-feature set. It cannot distinguish the effect of training against the reference anemometer from the remaining
140 difference between the two methods. The comparison is reproduced by `scripts/feature_ablation_analysis.py`.

S1.10 Model comparison setup

145 The comparison changes only the fitted machine-learning model. The feature matrix, folds, random seeds, metrics, and equal weighting of sessions remain unchanged. XGBoost v3.0.5 uses 1000 trees, maximum depth 7, learning rate 0.05, row and column subsampling of 0.8, squared-error loss, and early stopping after 50 rounds on a split by session within each training fold. LightGBM v4.6.0 uses 1000 trees, learning rate 0.05, all training rows, column subsampling of 0.8, and squared-error loss. None of these fixed settings was tuned on the validation data. The comparison is reproduced by `scripts/run_model_comparison.py`.

S1.11 Extended validation results: LOFO and L3FO

150 Table S4 compares the LOFO results, which hold out one complete session per fold, with the L3FO sensitivity analysis, which holds out three sessions at a time. Both schemes use the same feature matrix, model settings, and metric definitions. LOFO gives each session equal weight. Each L3FO metric pools the windows in its three held-out sessions before the combinations are averaged. The L3FO combinations overlap, so their spread describes sensitivity to the held-out set rather than an independent uncertainty.

Table S4. Ridge-model validation metrics under leave-one-flight-out (LOFO) and leave-three-flights-out (L3FO) cross-validation using 30 s windows. Values are mean $\pm$ standard deviation across sessions or combinations. Each L3FO value is first calculated from the pooled windows in a three-session combination. The combinations overlap, so their spread does not represent an independent uncertainty estimate. The $> 2 \text{ m s}^{-1}$ LOFO value retains sessions whose mean wind exceeds 2 m s^{-1} and averages their per-session direction MAE. The L3FO value retains combinations whose pooled mean wind exceeds the threshold. Speed MAE, component RMSE, and vector RMSE are reported for LOFO only.

	LOFO 50 folds	L3FO 19,600 comb.
Speed RMSE (m s^{-1})	0.32 ± 0.13	0.33 ± 0.10
Speed MAE (m s^{-1})	0.26 ± 0.11	–
Headwind-component RMSE (m s^{-1})	0.41 ± 0.21	–
Crosswind-component RMSE (m s^{-1})	0.36 ± 0.20	–
Vector RMSE (m s^{-1})	0.55 ± 0.28	–
Direction MAE ($^{\circ}$)	7.0 ± 4.9	7.0 ± 2.6
Direction MAE ($^{\circ}$), for $>2 \text{ m s}^{-1}$	6.0 ± 3.7	6.9 ± 2.5

S1.12 Additional model metrics

155 Table S5 reports speed MAE together with the headwind- and crosswind-component RMSEs across the 50 hover sessions. All values use the same 30 s windows and give each session the same weight as in the primary results.

Table S5. Additional metrics on the 50-session dataset using 30 s windows. Ridge values are from leave-one-flight-out cross-validation. Calibration-baseline values use fixed coefficients on the same windows. Speed errors are in m s^{-1} . Values are mean (standard deviation) across sessions. Bold values mark the lower error per row.

Metric	Ridge (LOFO)	Calibration baseline
Speed MAE	0.26 (0.11)	0.33 (0.13)
Headwind-component RMSE	0.41 (0.21)	0.47 (0.24)
Crosswind-component RMSE	0.36 (0.20)	0.46 (0.16)

S1.13 Selected feature definitions

Table S2 lists all 62 features. Table S6 defines the selected feature families used to interpret the two component models in Sect. 3.2 of the main article and gives the time basis of the lag and EWMA terms.

Table S6. Selected feature families and their definitions in the model.

Feature	Model definition
Mean pitch ($\bar{\theta}$)	Window-mean longitudinal tilt.
Mean roll ($\bar{\phi}$)	Window-mean lateral tilt.
Fast pitch/roll EWMA	Exponentially weighted history of retained window means with span 3.
Pitch/roll lag	Mean from one, two, or five retained windows earlier. These correspond to 1, 2, and 5 s when no QC gap intervenes.
Signed-squared tilt	Mean-square tilt retaining the sign of the corresponding mean angle.
Mean absolute tilt	Mean absolute pitch or roll within the window.

160 S1.14 Session with a relative lag

Session 2025-11-13 [13-27-22] has a much larger relative lag than the other sessions (Fig. S1). The whole-session cross-correlation lag has a magnitude of 38 s for the held-out Ridge trace. Because this diagnostic compares model and reference traces, it cannot distinguish a clock offset from other causes of a relative signal lag. The session speed RMSE is 1.1 m s⁻¹ for both Ridge and the calibration baseline. The session was retained in the primary evaluation, so the reported metrics include this lag. Excluding it changes the aggregate speed RMSE to 0.30 m s⁻¹ for Ridge and 0.38 m s⁻¹ for the calibration baseline while changing the direction MAE by less than 0.1°.

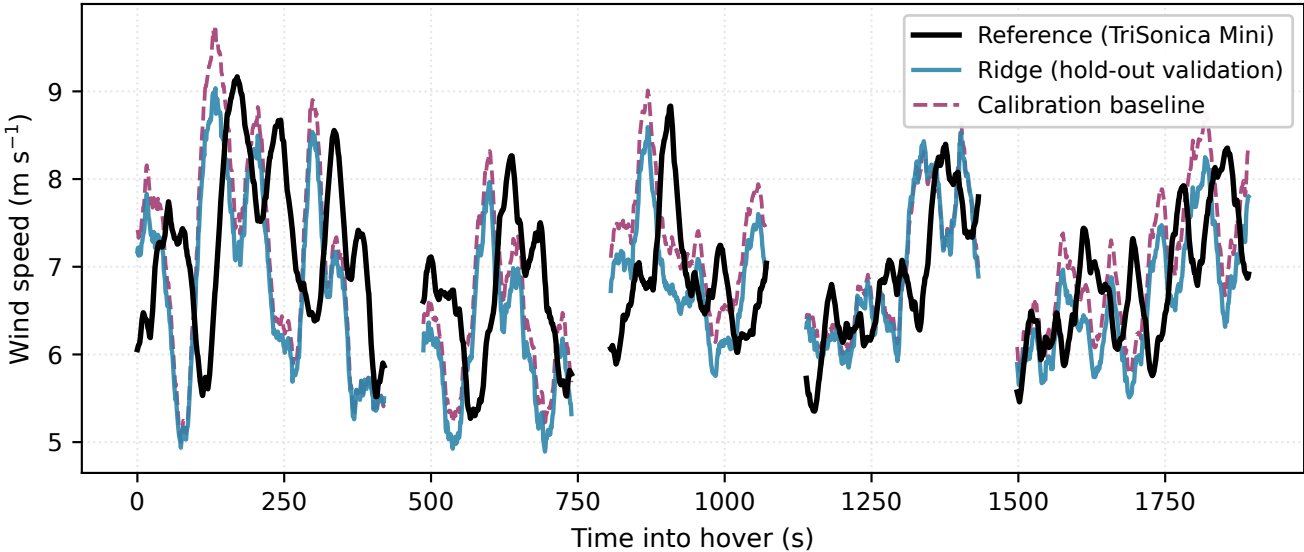


Figure S1. Wind-speed trace for session 2025-11-13 [13-27-22], showing the reference, held-out Ridge prediction, and calibration baseline. Gaps mark yaw-manoeuvre windows excluded by quality control. The telemetry-based estimates and reference trace exhibit a relative timing offset.

S1.15 Downwash check

A calm-day check found no detectable downwash effect beyond the stated sensor accuracy at a 2 m separation. At least this distance was maintained during the validation sessions, with the aircraft initially positioned downwind of the reference anemometer.

S1.16 Distribution of session-mean wind speeds

Figure S2 shows the distribution of session-mean wind speeds across the 50 hover sessions, categorized into the wind regimes of Table 3 of the main article. Most sessions have mean speeds below 5 m s⁻¹, and only six exceed 6 m s⁻¹. The low-, medium-, and high-wind results are therefore reported together with the number of sessions in each regime.

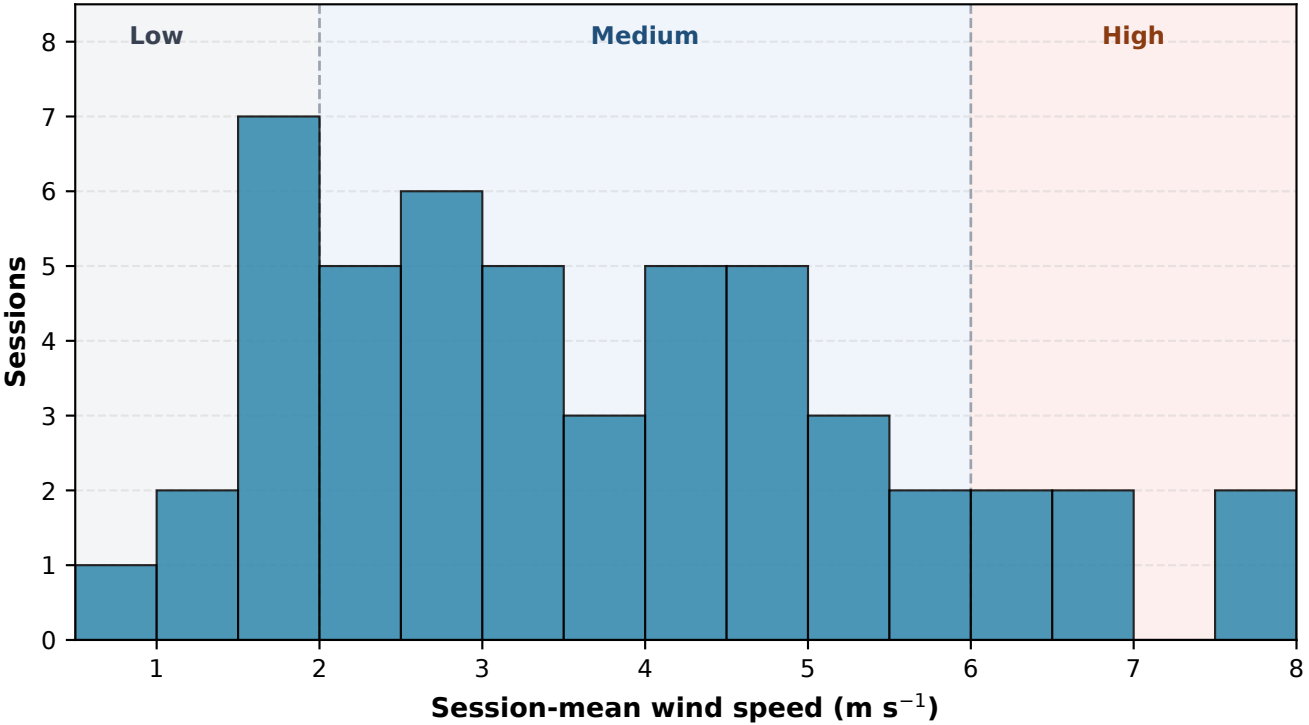


Figure S2. Distribution of session-mean wind speeds across the 50 validation sessions, with shaded backgrounds denoting the wind-speed regimes defined in Table 3 of the main article. Vertical dashed lines mark the 2 and 6 m s⁻¹ thresholds.

S1.17 Wind speed regime statistics

Table S7 divides the sampled conditions into seven narrower wind-speed intervals than the three broad regimes used in Table 3 of the main article. For each interval it reports the number of sessions, window-equivalent time, and the retrieval errors of both methods. The final row reproduces the aggregate metrics across all 50 sessions.

Table S7. Wind speed regime statistics. Window-equivalent time is the number of retained rows multiplied by the 1 s analysis step and does not represent unique elapsed time. Values are rounded to 0.1 h.

Regime	Sessions (n)	Window-equiv. time (h)	Ridge speed RMSE (m s ⁻¹)	Ridge direction MAE (°)	Baseline speed RMSE (m s ⁻¹)	Baseline direction MAE (°)
<1 m s ⁻¹	1	0.4	0.28	24.4	0.34	27.7
1–2 m s ⁻¹	9	3.6	0.32	9.5	0.41	12.3
2–3 m s ⁻¹	11	4.2	0.33	8.1	0.47	9.6
3–4 m s ⁻¹	8	3.3	0.25	5.6	0.33	7.1
4–5 m s ⁻¹	10	3.9	0.26	5.0	0.31	5.4
5–6 m s ⁻¹	5	1.9	0.30	3.8	0.37	5.1
>6 m s ⁻¹	6	2.6	0.49	6.5	0.48	6.5
Total	50	19.9	0.32	7.0	0.39	8.4

S1.18 High-wind session variability

180 Table S8 summarizes the six sessions exceeding 6 m s⁻¹. The relative-lag session has a speed RMSE of 1.09 m s⁻¹. Across the other five sessions, speed RMSE ranges from 0.27 to 0.44 m s⁻¹ and averages 0.37 m s⁻¹.

Table S8. Ridge (LOFO) performance for high-wind hover sessions (>6 m s⁻¹). Session-averaged values for all four models, including the calibration baseline, are given in Table S9. Excluding the relative-lag session, the remaining five sessions average a speed RMSE of 0.37 m s⁻¹ and a direction MAE of 6.4°.

Session	Mean wind (m s ⁻¹)	Speed RMSE (m s ⁻¹)	Direction MAE (°)
2025-11-18 [13-35-18]	6.44	0.267	7.2
2025-11-13 [14-02-34]	6.49	0.378	2.8
2025-11-13 [14-36-36]	6.65	0.377	3.7
2025-11-13 [13-27-22]	6.92	1.086	6.8
2025-11-18 [12-28-53]	7.58	0.389	10.6
2025-11-18 [13-02-32]	7.63	0.443	7.7

S1.19 Model families in the high-wind regime

185 The high-wind comparison uses the same held-out folds for all three machine-learning models and the same windows for the fixed calibration baseline. Table S9 summarizes the session-averaged performance across the six sessions above 6 m s⁻¹. With only six sessions above this threshold, the comparison is limited to describing the sampled cases.

Table S9. Session-averaged performance over the six high-wind hover sessions ($>6 \text{ m s}^{-1}$). The machine-learning models used the same leave-one-flight-out folds, and the fixed calibration baseline was evaluated on the same windows. Ridge and the calibration baseline have lower speed RMSEs than the tree ensembles, while direction MAE is similar across all four models. After excluding the relative-lag session described in Sect. S1.14, the average speed RMSEs across the other five sessions are 0.37 for Ridge, 0.50 for LightGBM, 0.55 for XGBoost, and 0.36 m s^{-1} for the calibration baseline. Their direction MAEs are 6.4, 6.2, 6.5, and 6.4° in the same order.

Model	Speed RMSE (m s^{-1})	Direction MAE ($^\circ$)
Ridge regression	0.49	6.5
LightGBM (gradient-boosted trees)	0.58	6.4
XGBoost (gradient-boosted trees)	0.63	6.7
Calibration baseline	0.48	6.5

S1.20 Diagonal wind error

The calibration baseline cannot represent cross-axis aerodynamic coupling and may therefore degrade under diagonal winds that tilt the aircraft on both axes. For winds above 2 m s^{-1} , the speed RMSE for diagonal winds exceeds that for cardinal winds by 0.01 m s^{-1} on average within sessions. The standard deviation across sessions is 0.14 m s^{-1} . The dataset does not resolve a directional coupling effect. The comparison is reproduced by `scripts/diagonal_error_analysis.py`.

S1.21 Paired session difference test

Subtracting the Ridge speed RMSE from the baseline value for each session gives $t = 5.79$ ($\text{df} = 49$, $p < 0.001$). A 10,000-resample bootstrap gives a 95 % confidence interval of $[0.051, 0.101] \text{ m s}^{-1}$. Sessions flown on the same day share weather, so a second bootstrap resampled the 15 measurement days and gave $[0.045, 0.108] \text{ m s}^{-1}$. For the 44 sessions with mean wind up to 6 m s^{-1} , the paired test gives $t = 6.55$ ($\text{df} = 43$, $p < 0.001$), and resampling the 14 represented days gives $[0.057, 0.120] \text{ m s}^{-1}$. All values are archived by `scripts/paired_session_test.py`.

S1.22 Leave-one-day-out sensitivity

The 50 hover sessions span 15 measurement days, with one to six sessions per day, so a leave-one-day-out (LODO) validation that holds out all sessions from one day prevents same-day training overlap (Roberts et al., 2017). The feature matrix, regularization, standardization, metrics, and equal weighting of sessions match LOFO throughout.

LODO gives a speed RMSE of $0.33 \pm 0.13 \text{ m s}^{-1}$ and a direction MAE of $7.0 \pm 4.5^\circ$. Under LOFO, the corresponding values are $0.32 \pm 0.13 \text{ m s}^{-1}$ and $7.0 \pm 4.9^\circ$, so the changes amount to 0.009 m s^{-1} and less than 0.1° , respectively. The check is reproduced by running `scripts/run_ridge_robustness.py` with the `--lodo` flag.

S1.23 Gustiness and the residual high-wind error

The six sessions above 6 m s^{-1} contain 9,285 windows. Absolute Ridge speed error has a correlation of $r = -0.065$ with the variability of reference speed within each window. The partial correlation is $r = -0.147$ after controlling for window-mean speed. Mean absolute errors by variability quartile are 0.40, 0.45, 0.39, and 0.34 m s^{-1} from lowest to highest. These window-level associations are descriptive because adjacent windows overlap. The sampled windows do not show larger errors under greater horizontal variability. Effects of vertical motion cannot be tested because the aligned products contain neither aircraft altitude nor vertical wind (Sect. S1.1). The analysis is reproduced by `scripts/gustiness_analysis.py` and writes `data/results/gustiness_analysis.csv`.

References

- 215 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.,
Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python,
Journal of Machine Learning Research, 12, 2825–2830, 2011.
- Polle, A.: Trisonica Datalogger for Raspberry Pi, https://github.com/a-polle/Trisonica_Datalogger_RaspberryPi, last access: 19 March 2026,
2026.
- 220 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller,
W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical,
or phylogenetic structure, *Ecography*, 40, 913–929, <https://doi.org/10.1111/ecog.02881>, 2017.
- Vauvillier, L.: dji-log-parser, <https://github.com/lvauvillier/dji-log-parser>, last access: 17 December 2025, 2025.