IN-FIELD CALIBRATION OF UAV THERMAL IMAGERY FOR PRECISION AGRICULTURE

by

Sheamus Vaughan

A professional paper submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

In

Land Resources and Environmental Sciences

MONTANA STATE UNIVERSITY
Bozeman, Montana

May 2026

©COPYRIGHT
by

Sheamus Vaughan

2026

All Rights Reserved

TABLE OF CONTENTS

1. INTRODUCTION 1

Hypotheses 3

2. METHODS 4

Data Colletcion and Proccessing 4
Post-deplyment Calibration Analysis 6

3. RESULTS 9

Tables and Figures 15

4. DISCUSSION 18

5. LIMITATIONS 22

6. REFERENCES 24

LIST OF TABLES

Tables Page

1. Table 1. Statistical metrics for different calibration methods. 14

2. Table 2. Statistical metrics (Generalized ELM with FPA) for different environmental conditions 15

LIST OF FIGURES

Figures Page

1. Figure 1. RMSE for all seven calibration methods. 15
2. Figure 2. Factory vs. best fit calibration method with 1:1 line between UAV and radiometer data. 16
3. Figure 3. Temperature residuals with and without FPA for the Temporal drift model 16
4. Figure 4. Raw and calibrated (model 3B) thermal mosaics, masked to >.3 NDVI 17
5. Figure 5. Relative difference in CWSI between raw and calibrated mosaics 17
6. Figure 6. Distributions of pixel temperatures for raw and calibrated mosaics, includes published envelopes for barley 18

ABSTRACT

My study investigates in-field calibration of drone based thermal imagery as a precision tool for assessing water stress. High-resolution canopy temperature mapping serves as a critical indicator of plant water use dynamics. Collecting actionable measurements requires accurate sensor calibration to overcome bias in unmanned aerial vehicle (UAV) mounted uncooled long-wave infrared (LWIR) cameras. I obtained reference temperature measurements using two high-emissivity foam panels monitored by NIST-traceable Apogee SI-121 infrared radiometers. Based on these target temperatures I evaluated seven calibration method variants on 106 synchronized radiometer-sensor pairs from a single flight over a barley (*Hordeum vulgare*) field located near Bozeman, Montana. I compared the factory baseline against calibration models using empirical line method (ELM) with temporal drift correction and generalized versions, both incorporating the camera's focal plane array (FPA) temperature as a covariate. RANSAC (RANdom SAmple Consensus) robust regressions minimized the influence of outliers during model fitting, with Monte Carlo cross-validation (200 iterations) and bootstrap resampling (1,000 iterations) used to quantify uncertainty. The factory calibration produced an RMSE of 4.11°C, with opposite-sign gain errors of approximately +4.3°C on the cold panel and -2.5°C on the hot panel. Post-deployment calibration reduced RMSE by up to 44.8%, with the best-performing method, generalized LWIR+FPA, achieving RMSE=2.27°C. Applying this calibration to the full flight mosaic shifted mean canopy temperature from 39.97°C to 32.42°C and produced agronomically plausible Crop Water Stress Index values consistent with late-season dry-farmed barley under high vapor pressure deficit; whereas the uncalibrated imagery placed the entire canopy distribution implausibly above the non-transpiring upper limit. These results demonstrate that incorporating brief reference target imaging at takeoff and landing substantially improves the accuracy of UAV thermal data products for crop water stress assessment.

1. Introduction

Accurate maps of surface temperature derived from unmanned aerial vehicle (UAV) thermal infrared imaging (also called thermography) is essential for estimating crop water use and energy budgets, as well as for validating satellite-derived data (Ribeiro-Gomes et al., 2017; Messina & Modica, 2020; Niu et al., 2020). Even small errors in estimates of temperature from longwave radiation propagate into substantial errors in evapotranspiration metrics and crop water indices (Zhao et al., 2018; Das et al., 2021; Park et al., 2021). While the underlying radiance-temperature relationships implemented in modern long-wave infrared (LWIR; the spectral region in which surfaces emit thermal radiation) sensors are well established, field operating conditions introduce additional uncertainty that is not fully captured by the factory lookup-table conversions (Kelly et al., 2019). Rigorous post-deployment calibration, linking sensor-reported temperatures to independently measured reference temperatures under field conditions, provides greater confidence in resulting data products for research, agricultural management, and irrigation decision making (Zhao et al., 2018; Das et al., 2021; Han et al., 2020). This particular work has been motivated by the need to quantify temperature between barley (*Hordeum vulgare*) varieties with and without-stay green traits (Williams et al., 2022).

My research aims to establish whether calibration of UAV thermal imagery against high-emissivity field reference panels, whose surface temperatures are independently measured by co-located National Institute of Standards and Technology (NIST; the U.S. federal authority that maintains instrument reference standards)-traceable infrared radiometers, provides operationally meaningful improvements in accuracy over manufacturer-provided factory calibrations. In published literature, factory calibrations typically exhibit substantial errors, with mean absolute error (MAE) often approaching 5°C (Kelly et al., 2019; Gerchow et al., 2025). A dual-target

design using panels at two different temperatures can reveal gain (slope error): a systematic miscalibration that produces errors of opposite sign at different target temperatures, missing in single point reference (Lin et al., 2022; Kelly et al., 2019; Zhao et al., 2018). Because the opposite-sign errors partially cancel when both panels are pooled into a single RMSE, a dual-panel evaluation produces a narrower baseline relative to single-panel studies (Kelly et al., 2019; Zhao et al., 2018). The two-point calibration lowers both the starting RMSE as well as the achievable percentage reduction, imposing a more demanding test of residual gain error across the full thermal range (Lin et al., 2022).

The goal of exploring alternative calibration models is to create a procedure for systematically improving thermal imagery quality in UAV-based crop monitoring. The key differences among calibration approaches lie in how training data are selected and whether focal plane array (FPA; the grid of detector elements at the camera's focal plane) temperature is included as an additional predictor. To test hypothesis 1, I compared the factory baseline to three post-deployment calibration approaches: Two temporal methods, with both static and interpolated versions, in which training data are drawn from the beginning and end of the flight, and a generalized method, which draws training data randomly from across all flight phases. I implemented both methods in two variants: one including FPA temperature as an explicit predictor and one omitting it (Aragon et al., 2020). For the temporal, the static and interpolated sub-variants each have FPA and no-FPA versions, creating the four temporal method variants (static LWIR, static LWIR+FPA, interpolated LWIR, interpolated LWIR+FPA). This paired comparison directly tests whether parameterizing FPA temperature materially improves calibration beyond what the empirical line method captures through implicit temporal structure.

The temporal calibrations test whether training primarily on the hovering periods at the beginning and end of flights improves prediction of mid-flight measurements, addressing the temporal drift question posed in hypothesis 2. The generalized models test whether models trained on a randomly selected set of data generalize effectively to unseen segments without additional temporal context, considered in hypothesis 3. Using this framework I compared the four calibration strategy types (factory, temporal 1 & 2, and generalized), these comprise seven variants in total. I implemented each approach with and without FPA temperature as a covariate, serving as hypothesis 4. The combination of these different conditions yields the seven methodologies tested.

Hypothesis 1 (Factory vs. Post-Deployment): Bias in temperature inferred from thermal infrared irradiance can be corrected relative to the factory calibration because uncooled thermal sensors experience predictable systematic temperature-dependent drift relative to reference data and are often operated under ambient conditions that systematically differ from factory calibration environments.

Hypothesis 2 (Temporal Drift): Using reference data from the beginning and end flight segments will provide improved calibration of mid-flight temperatures because the relationship between sensor output and true temperature is expected to shift systematically over the course of flight operations.

Hypothesis 3 (Generalization): Calibrations developed from data spanning a broad range of sensor and environmental states will generalize effectively to unseen test data. Avoiding inaccurate extrapolation requires training data that represents the full operational envelope, which can be achieved using Monte Carlo cross validation. A calibration developed under narrow

conditions may fail when applied to flight data where the FPA reaches temperatures outside the training range.

Hypothesis 4 (FPA-Driven Error): Including FPA temperature as an explicit predictor variable will further reduce bias and MAE compared with time-based corrections alone because sensor drift is mechanically linked to the thermal state of the focal plane array rather than simply elapsed time.

2. Methods

2.1 Data collection and processing

Thermal infrared imagery collection was conducted collaboratively with the MSU barely program at Post Agronomy Farm in Bozeman (~45.40° N, 111.09° W; ~1,460 m elevation), Montana evaluated on August 06th 2025. We used a DJI Matrice 350 RTK UAV to acquire data over a barley (*Hordeum vulgare)* field between 1100–1600 hours local time, at altitudes of 60–80 m, ground speeds of approximately 8 mph, and with 75% imagery overlap; configurations which correspond with standard mapping operations. We equipped the UAV with an AgEagle Altum-PT multispectral camera that uses a radiometrically calibrated FLIR Boson microbolometer for longwave infrared (LWIR) imaging (7.5–13.5 µm spectral bandwidth; 320 × 256 resolution). Reference target observations were situated during the beginning, middle, and end of flight operation (Han et al., 2020).

In a microbolometer, scene temperature is inferred by the change in electrical resistance of each detector element, which depends on the temperature difference between the absorber and the FPA; brightness temperature reported by the camera is therefore sensitive to the camera's own internal temperature state (Kelly et al., 2019). Because UAV deployment exposes the camera to

changing ambient conditions, in which the FPA drifts through a flight, the Boson firmware periodically performs a shutter-based flat-field correction (FFC), in which an internal shutter of approximately uniform, known temperature is interposed between the optics and the FPA to re-zero pixel offsets. Kelly et al. (2019) showed that for an uncalibrated UAV-mounted thermal camera, this internal correction was not sufficient to remove flight-induced drift, motivating the dual-panel external calibration applied here.

Two high-emissivity foam calibration panels serve as extended-area thermal targets, with their surface temperatures independently measured by co-located Apogee SI-121 infrared radiometers (designated TC1 over the cold panel and TC2 over the hot panel) to provide the calibration reference standard. Each radiometer mounted above its respective target and oriented vertically downward with a 35° field of view encompassing the panel surface (Zhao et al., 2018; Aragon et al., 2020). These radiometers provide spatially averaged surface temperature measurements that serve as the reference standard for calibration. The radiometers have factory calibrations traceable to NIST standards and an absolute accuracy of ±0.2°C (Aragon et al., 2020). During the flight on 8/6, TC1 (cold) maintained a mean target temperature of 41.4°C and TC2 (hot) maintained 59.8°C, providing an ~18°C reference temperature spread that enables evaluation of sensor accuracy (by including gain/slope error) across the thermal dynamic range, rather than at a single reference point.

Image segmentation performed in Roboflow identifies frames from the UAV-acquired thermal imagery containing the foam calibration targets. For each identified frame, a region-of-interest (ROI) mean temperature is calculated for the pixels covering the foam panels (Paciolla et al., 2025). I then synchronized these UAV-derived temperatures by timestamp with corresponding ground-reference radiometer measurements to create a dataset of 106 synchronized radiometer-

sensor pairs (53 cold-panel and 53 hot-panel detections). Each observation passes through a five-stage pipeline: **(1)** image capture and raw radiometric extraction, **(2)** panel detection and ROI temperature computation in Roboflow, **(3)** timestamp synchronization with the radiometer datalogger, **(4)** temporal segment assignment, and **(5)** model fitting and cross-validation.

My analysis of the thermal imagery contains an assumption regarding emissivity. Sensor and radiometer reported values are treated as apparent (brightness) temperatures under an emissivity of $\varepsilon=1$, where emissivity is the ratio of radiant energy emitted by a surface to that emitted by an ideal blackbody at the same temperature (Messina & Modica, 2020, Sobrino et al., 2008). This is appropriate for the calibration procedure itself, since both instruments view the same high-emissivity panel surface under similar geometry. The assumption is more consequential when calibrated temperatures are used to retrieve canopy values, where leaf emissivity (~0.97) and reflected clear sky downwelling radiance introduce a small cool bias in uncorrected brightness temperatures (Sobrino et al., 2008). A post-mosaic emissivity and sky-radiance correction is recommended for greater accuracy of water-stress, but is outside the present scope of my sensor calibration analysis.

2.2. Post-deployment calibration analysis

I originally implemented the calibration analysis in RStudio and subsequently migrated it to a Python-based analysis pipeline using Claude Code, preserving the original methodology and output structure. All calibration methods use the same fundamental model structure: an empirical line method (ELM) that fits a linear relationship between UAV-measured temperatures and radiometer reference temperatures (Gerchow et al., 2025). Since thermal imagery can contain anomalous measurements from transient events (e.g., passing clouds, momentary vibrations), a robust fitting approach is needed. The Random Sample Consensus (RANSAC) algorithm

addresses this by repeatedly fitting candidate models to randomly selected small subsets of the data, then selecting the solution that consistently explains the largest number of measurements within a specified tolerance (Fischler et al., 1981). This approach minimizes the influence of outliers on the final calibration coefficients. I set the inlier threshold as the median absolute deviation of residuals from an initial least-squares fit.

For the temporally interpolated model the repeated empirical line method (RELM) fits a linear model of the form $T_Cal = \beta_0 + \beta_1 \cdot T_Raw + \beta_2 \cdot T_FPA$, where T_Raw is the UAV-reported temperature, T_FPA is the optional focal plane array temperature, and β_0, β_1, and β_2 are fitted coefficients. The method is "repeated" because a separate linear calibration model is fit independently for each flight phase anchor point (begin and end), and predictions for intermediate observations are derived by interpolating between those independently fitted models, as opposed to a single model fit to all data at once. I estimated parameters via RANSAC-robust regression on synchronized radiometer reference pairs, with temporal interpolation applied across flight phases for the repeated variant. Because all variants use the same underlying linear model equation, differences in performance can be attributed to how training data are selected and how sensor drift is handled, not to differences in model form (Gerchow et al., 2025).

Method 1: Factory baseline. Unmodified camera readings establish the baseline error benchmark against which I evaluated all subsequent calibration methods. Published research indicates factory calibrations typically produce MAE of 2–5°C due to uncorrected environmental effects and temporal drift (Zhao et al., 2018; Kelly et al., 2019). Because this study employs a dual-panel design, the pooled baseline RMSE reflects partial cancellation of opposite-sign gain errors across the two target temperatures; the reported factory baseline is therefore lower than

would be expected from a single-panel evaluation, the reader should interpret percentage improvements across methods accordingly (Lin et al., 2022; Kelly et al., 2019).

Method 2: Temporal method. This approach addresses anticipated sensor drift by training only on temperature pairs from the beginning and end of each flight, on the assumption that these periods represent more stable thermal states than mid-flight (Kelly et al., 2019). I assigned temporal segments algorithmically based on elapsed flight duration: First the algorithm executes a time-based cut, assigning observations chronologically to the begin and end segments based on the first and second largest gaps in elapsed flight time, with all remaining observations forming the mid segment. Based on our flight data (approximately 33 minutes of panel overflight time), this yielded partitions: begin N=34, mid N=26, and end N=46. All begin- and end-segment observations serve as training data (approximately 75% of total); mid-segment observations form the test set (approximately 25%). The static variant fits the single RANSAC model on the combined begin and end data and applies it uniformly to mid-flight predictions. While the interpolated variant uses the mean timestamp of each segment as the anchor, so the coefficient at any observation reflects its proportional elapsed-time position between the begin and end segment means, addressing the possibility that sensor drift is non-linear across the flight.

Method 3: Generalized method. This approach uses random 80–20 partitioning across all timestamps regardless of flight phase (Gerchow et al., 2025). Monte Carlo cross-validation with 200 random splits constrains the distribution of error correction performance that might be expected using this approach. Training data (80%) fit model equations and test sets (20%) provide unbiased performance assessment. I did not use a separate validation set; RANSAC provides internal robust fitting during the training phase.

Primary metrics used in assessing model performance include Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) of the post-deployment calibrated temperatures relative to the held-out test sets that are assumed to represent the accurate temperatures of the foam reference panels (Han et al., 2020; Han et al., 2021). I quantified model bias as the mean signed difference between predicted and reference temperatures (positive values indicate systematic overestimation). I also computed the coefficient of determination (R^2). I quantified uncertainty in the generalized ELM via Monte Carlo cross-validation (stable at 200 iterations, 80/20 splits), which characterizes variability in test-set performance across random data partitions. I assessed temporal RELM stability via bootstrap resampling (1,000 iterations) of the begin and end training segments, providing 95% confidence intervals on the temporal model's sensitivity to specific observations in training. This differs from the Monte Carlo intervals which reflect within-partition resampling uncertainty rather than partition-selection uncertainty.

3. Results

Post-deployment calibration reduced error relative to the factory baseline by as much as 44.8%, with the best-performing method (Generalized LWIR+FPA, Method 3B) achieving RMSE=2.27°C, consistent with the magnitude of improvement reported in peer-reviewed studies (Aragon et al., 2020; Han et al., 2020). Across all methods, I computed performance metrics on a dataset of N=106 synchronized radiometer-sensor pairs from the flight, comprising 53 cold-panel and 53 hot-panel detections spanning a mean panel temperature spread of approximately 18°C (TC1 cold: 41.4°C; TC2 hot: 59.8°C). Temporal partitioning used time-based segment assignment: begin N=34, mid N=26, end N=46 (Figure 2).

The factory specifications produced an RMSE of 4.11°C and MAE of 3.61°C on the full dataset (Table 1), with a positive pooled bias of +0.92°C (Aragon et al., 2020). Evaluated per target, the

factory overestimated the cold panel (TC1) by approximately +4.3°C on average and underestimated the hot panel (TC2) by approximately -2.5°C, opposite-sign gain errors that pooled to produce the RMSE of 4.11°C. These values are consistent with pre-calibration error magnitudes reported in comparable UAV thermal studies (Aragon et al., 2020; Han et al., 2020; Kelly et al., 2019), where uncalibrated RMSE values typically range from 3.4°C to over 14°C. Temporal ELM calibrations reduced test RMSE to 2.82°C without FPA temperature (static, 2A) and 2.60°C with FPA (static, 2B) (Figure 1), corresponding to RMSE reductions of 31.3% and 36.8% relative to the baseline (Table 1). The interpolated temporal variant without FPA (2C) achieved RMSE=3.43°C (16.4% reduction); the FPA-augmented interpolated variant (2D) performed worse than the factory baseline at RMSE=5.39°C and is not a recommended method. Mean absolute error for the static temporal models was 2.16°C (2A) and 2.04°C (2B). The static temporal models (2A, 2B) achieved R^2 of 0.918 and 0.930 on the mid-flight test set, indicating that a single begin+end model generalizes reasonably across mid-flight observations. The interpolated temporal variant (2C) achieved R^2=0.878, reflecting greater sensitivity of the coefficient to the specific observations in each segment. The factory calibration achieved R^2=0.808, confirming that the camera's radiometric response is approximately linear but is systematically offset by gain error (Figure 2).

Monte Carlo cross-validation with 200 random 80–20 splits quantified uncertainty in the generalized ELM performance. The generalized model without FPA achieved a mean test RMSE of 2.51°C (MAE 1.98°C, bias +0.14°C, R^2 0.927) across the 200 realizations, while the FPA-augmented generalized model had a mean test RMSE of 2.27°C (MAE 1.73°C, bias -0.19°C, R^2 0.940) (Figure 1), representing 38.9% and 44.8% RMSE reductions relative to the factory calibration (Table 1). The 95% Monte Carlo interval for the non-FPA model spanned 1.91–

3.21°C and for the FPA model spanned 1.53–3.10°C, indicating that performance is relatively stable across data partitions but can vary by approximately ±0.8°C depending on which observations are drawn into training.

I used bootstrap resampling (1,000 iterations) of the temporal training segments to assess the stability of temporal ELM coefficients to the specific observations in training. The 95% bootstrap stability interval for the static temporal model without FPA (2A) spanned 2.66–3.20°C, while the FPA-augmented static model (2B) was more constrained at 2.57–2.78°C, indicating that FPA temperature absorbs drift-related variance and produces more stable coefficients. Bootstrap analysis confirmed that including FPA reduced RMSE in 99.4% of resampled iterations for the static temporal model, with a mean reduction of 0.26°C (95% CI: 0.046–0.542°C). For the interpolated models, FPA inclusion was harmful in 97.9% of iterations (mean increase 2.04°C), consistent with the 2D method's substantially worse observed performance.

The best-performing method overall, Generalized LWIR+FPA (Method 3B), achieved R^2=0.940 and RMSE=2.27°C on held-out test observations, outperforming the factory baseline (R^2=0.808, RMSE=4.11°C) and the static temporal models (2A: R^2=0.918; 2B: R^2=0.930) (Table 1). The generalized approach's advantage over the temporal approach likely reflects that training on observations drawn from across the full flight provides better coverage of the FPA temperature range encountered during mid-flight, avoiding the extrapolation risk that arises when the mid-flight FPA state lies outside the beginning-end training envelope.

The variation of calibration error across wind speed and air temperature bins (Table 2) does not constitute evidence that environmental conditions are meaningful drivers of calibration error. With N=106 observations divided across three bins per variable (N=22–48 per bin), the stratified analysis lacks the statistical power to distinguish real environmental effects from random

partition variance. The wind-speed pattern, lowest RMSE under high-wind conditions (1.04°C) and highest under medium-wind conditions (2.96°C), is non-monotonic and inconsistent with a simple physical mechanism; if convective effects were the primary driver, errors would be expected to increase or decrease directionally with wind speed rather than peak at intermediate values. Air temperature showed a similarly irregular pattern, with moderate air temperatures producing lower errors than either cool or warm conditions. These non-monotonic, cross-variable patterns are more consistent with random sampling variability across a single flight than with a systematic environmental signal. I did not conduct a formal test of environmental significance, the results of Table 2 should be read as exploratory and descriptive only. Multi-flight datasets spanning genuinely different meteorological conditions will be required before environmental factors can be identified as statistically or operationally significant contributors to calibration error (Zhao et al., 2018; Kelly et al., 2019).

Applying the best-performing calibration (Method 3B, Generalized LWIR+FPA) to the full flight thermal mosaic changed both canopy temperatures (Figure 4) and crop water stress estimates (Figure 5) compared with the uncalibrated raw mosaic. I identified crop areas with an NDVI $\geq$ 0.3 mask co-registered with the optical mosaic, then applied a 2% trim to the canopy min and max temperature distribution, leaving 1,617,312 canopy pixels in analysis. Mean canopy surface temperature dropped from 39.97°C in the raw imagery to 32.42°C after calibration, an offset of -7.55°C (MAE = 7.66°C, RMSE = 8.48°C) (Figure 4). The Method 3B model applied to every canopy pixel was $T_{cal} = 26.60 + 1.56 \cdot T_{raw} - 1.51 \cdot T_{FPA}$. The LWIR slope of 1.56 means that a 1°C change in raw signal corresponds to a 1.56°C change in calibrated temperature. This serves to expand the dynamic range that the raw sensor compresses. The negative FPA coefficient (-1.51) shows that as the focal plane array warms during flight and calibration

removes that drift. Pixel-level agreement between raw and calibrated temperatures was strong (Pearson r = 0.91). Calibration also widened the canopy temperature distribution (SD 3.47°C raw vs. 5.26°C calibrated) (Figure 4), preserving relative spatial structure while expanding the absolute thermal range (Figure 6).

Translating these temperatures into crop water stress requires agronomic research from Idso (1982) on post-heading spring-barley non-water-stressed baseline (lower limit LL = 25.65°C at the flight VPD of 2.53 kPa) and the simplified non-transpiring upper limit UL = Tair + 5°C = 32.03°C (Idso et al., 1981), which together produced noticeably different CWSI values for the two mosaics (Figure 5). Figure 6 expresses the raw and calibrated canopy distributions as Tcanopy - LL in degrees C, directly comparable to literature reports of stressed cereal canopy temperatures (Kemanian et al., 2005). Mean raw canopy temperature sat 14.39°C above the well-watered baseline (SD 3.47°C), roughly 13°C above air temperature, far beyond the 2–4°C above-air range reported for water-stressed cereals (Kemanian et al., 2005) and the +5°C non-transpiring upper limit (Idso et al., 1981). After calibration, the canopy distribution centered at +6.94°C above LL (SD 5.26°C), near the non-transpiring upper limit, with a substantial portion of the distribution falling within the literature stressed-cereal range. Normalizing these temperatures to the conventional simplified CWSI = (Tcanopy - LL) / (UL - LL) places a well-watered canopy at 0 and a non-transpiring canopy at 1 (Paciolla et al., 2025). The mean raw was 2.25 (SD 0.54) versus 1.09 (SD 0.83) for calibrated. 45.5% of the calibrated canopy was above CWSI = 1 and 3.8% was below 0, consistent with late-season dry-farmed barley under high VPD; the raw distribution lay entirely above CWSI = 1 (Figure 6).

Table 1) Statistical metrics for different calibration methods

Method	RMSE (°C)	MAE (°C)	Bias (°C)	R^2	Reduction (%)	95% CI (°C)
Factory Baseline (1)	4.11	3.61	+0.92	0.808	-	-
Temporal Static LWIR (2A)	2.82	2.16	-0.88	0.918	31.3%	2.66–3.20[1]
Temporal Static LWIR+FPA (2B)	2.60	2.04	-0.23	0.930	36.8%	2.57–2.78[1]
Temporal Interp LWIR (2C)	3.43	2.77	-2.21	0.878	16.4%	3.05–3.78[1]
Temporal Interp LWIR+FPA (2D)	5.39	4.83	-4.68	0.699	-31.2%	3.51–7.88[1]
Generalized LWIR, MC Mean (3A)	2.51	1.98	+0.14	0.927	38.9%	1.91–3.21[2]
Generalized LWIR+FPA, MC Mean (3B)	2.27	1.73	-0.19	0.940	44.8%	1.53–3.10[2]

[1] Bootstrap 95% stability CI (1,000 iterations) [2] Monte Carlo 95% CI (200 iteration)

Table 2) Statistical metrics (Generalized ELM with FPA) for different environmental conditions

Condition	N	MAE (°C)	RMSE (°C)	Bias (°C)
Wind Speed				
Low Wind	36	1.66	2.09	+0.53
Medium Wind	38	2.46	2.96	-0.41
High Wind	32	0.77	1.04	+0.10
Air Temperature				
Cool Air	36	2.20	2.77	-0.92
Moderate Air	48	1.20	1.69	+0.19
Warm Air	22	1.86	2.23	+1.42
Method: Generalized LWIR+FPA (3B). Single-flight dataset; results are exploratory.				

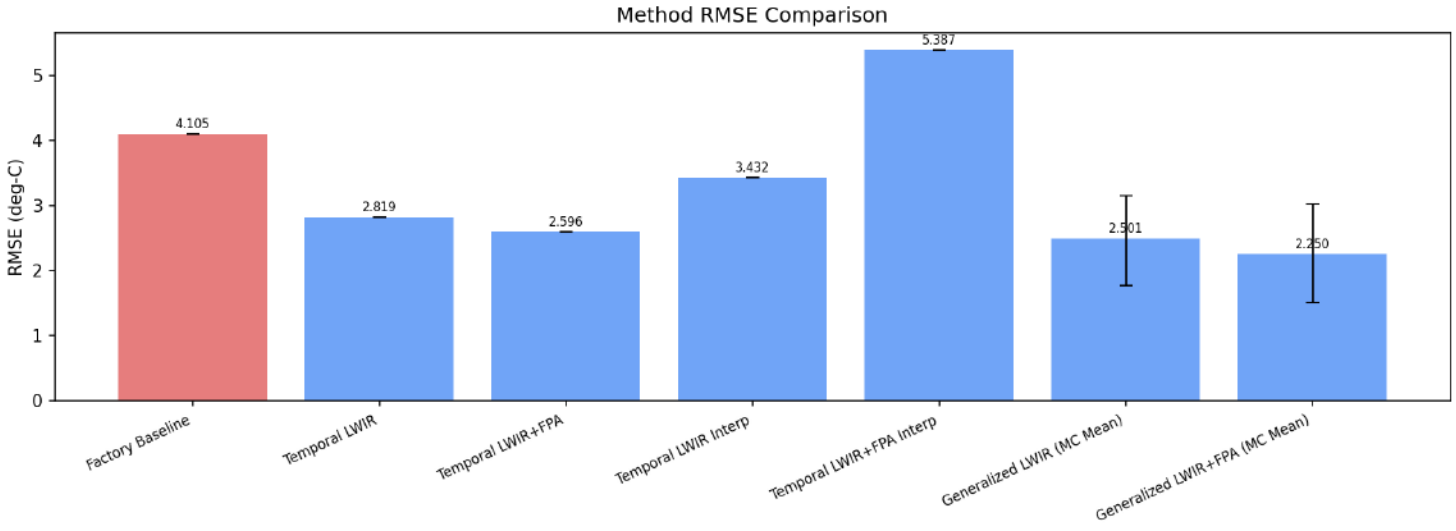


Figure 1) RMSE for all seven calibration methods.

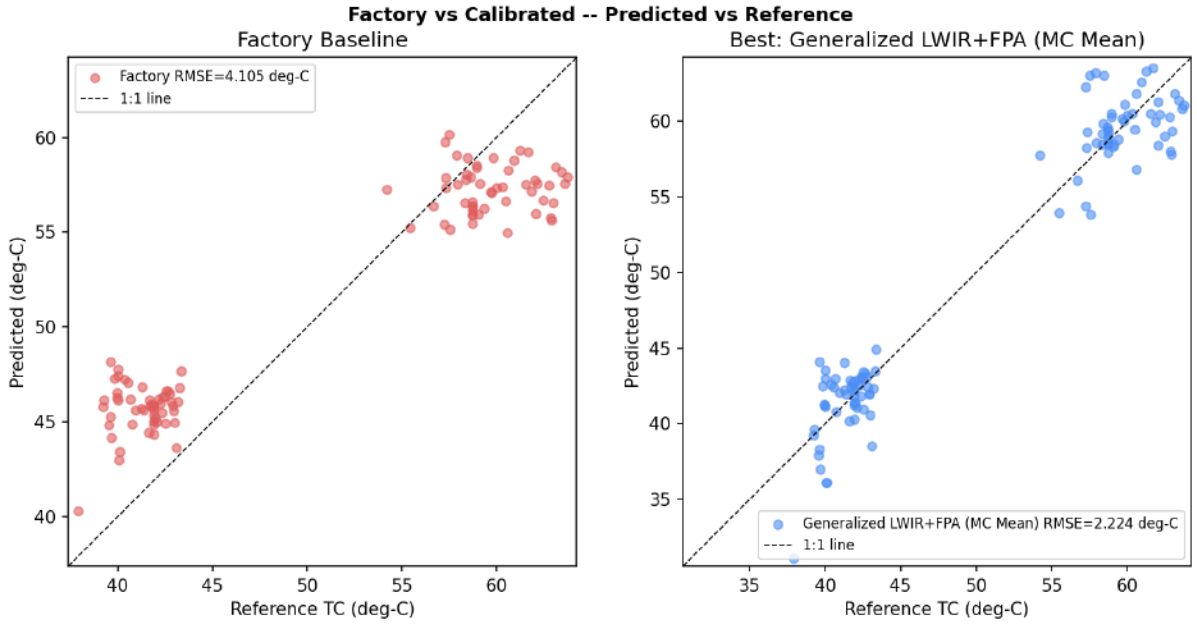


Figure 2) Factory vs. best fit calibration method with 1:1 line between UAV and radiometer data.

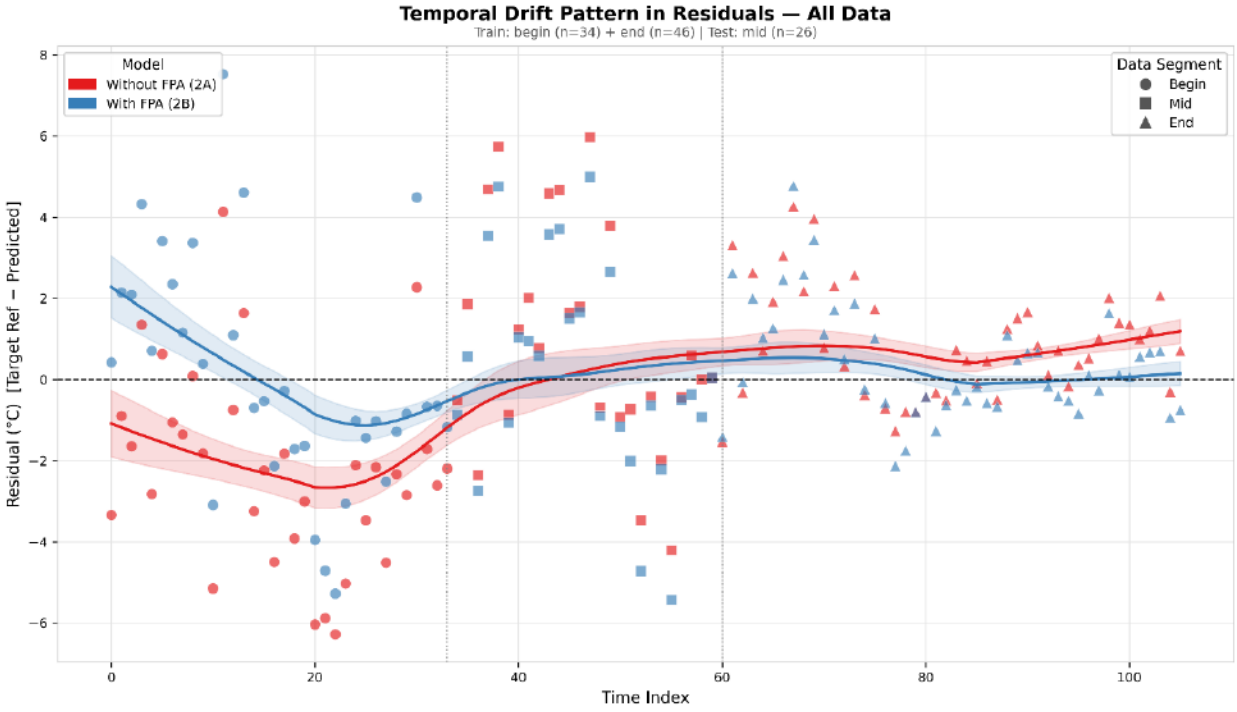


Figure 3) Temperature residuals with and without FPA for the Temporal drift model.

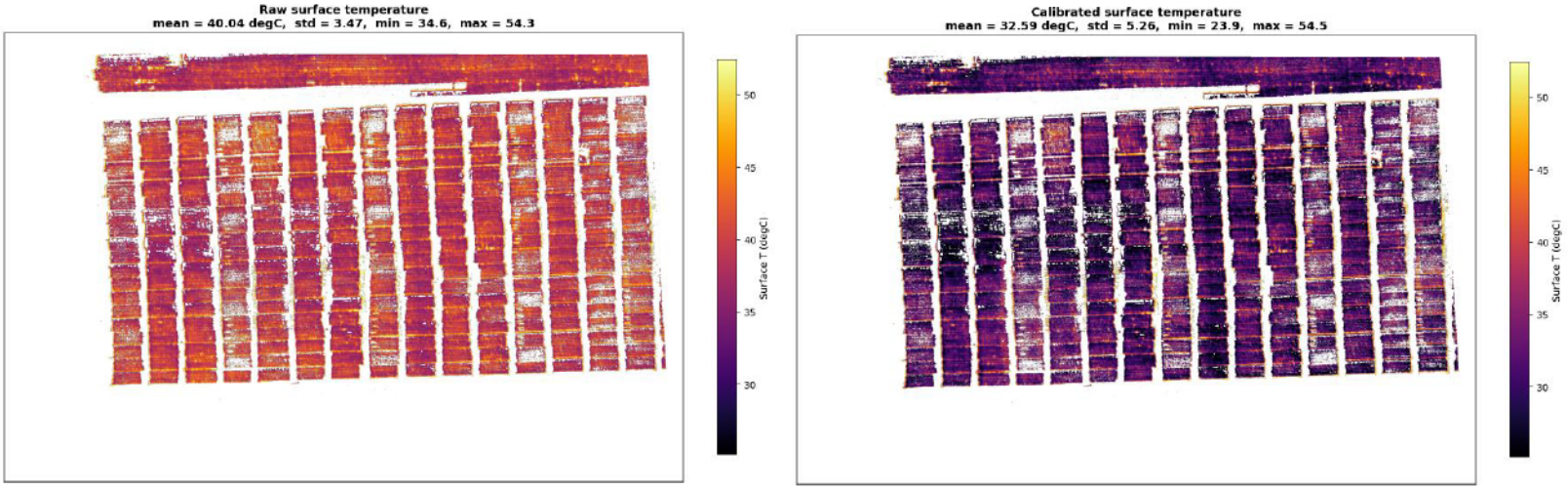


Figure 4) Raw and calibrated (model 3B) thermal mosaics, masked to >.3 NDVI.

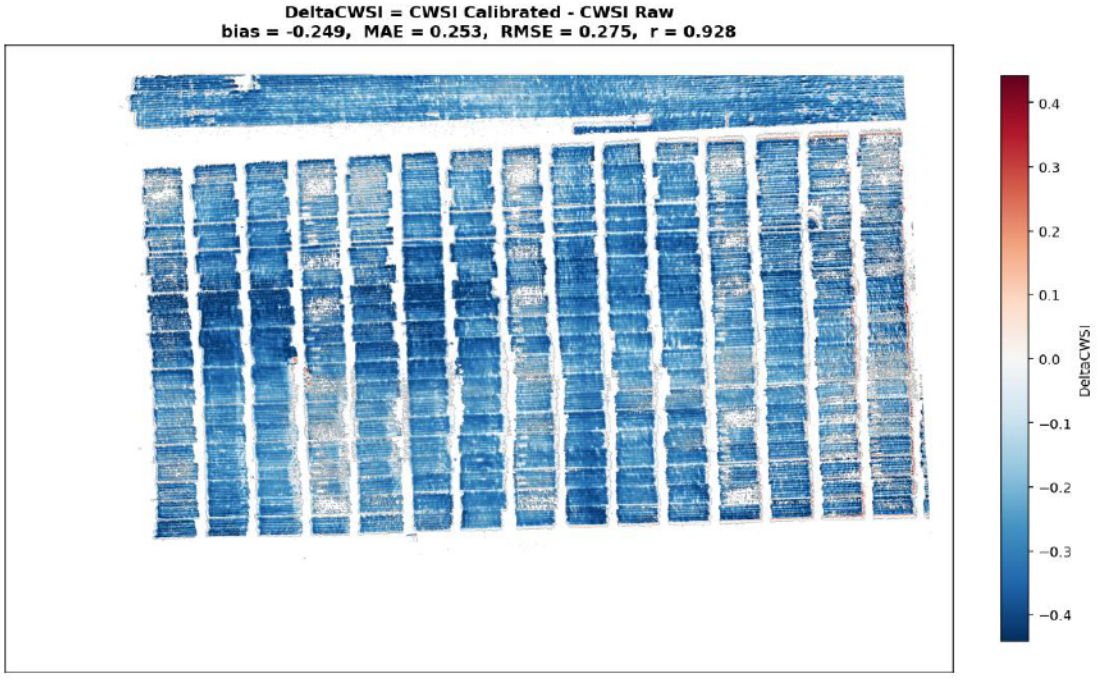


Figure 5) Relative difference in CWSI between raw and calibrated mosaics.

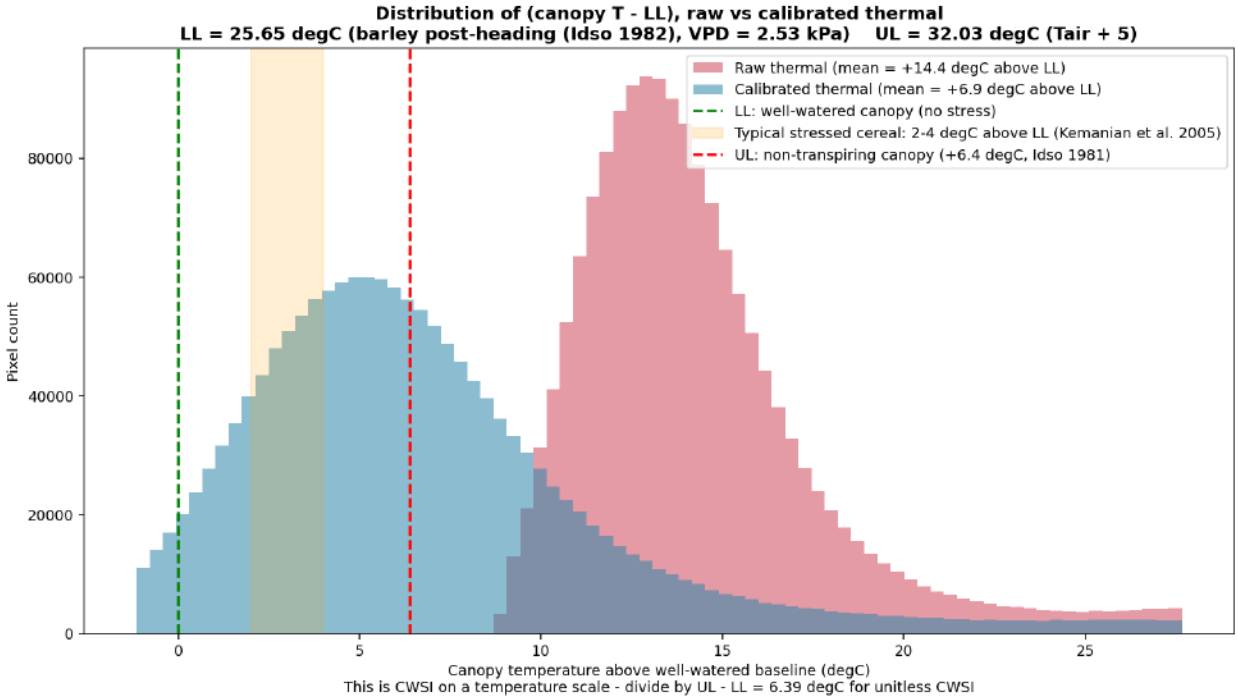


Figure 6) Distributions of pixel temperatures for raw and calibrated mosaics, includes published envelopes for barley.

4. Discussion

My results demonstrate that post-deployment calibration using two-point reference targets provides operationally meaningful improvement over factory calibration (Figure 1), confirming Hypothesis 1. The factory calibration exhibited per-target gain error, it overestimated the cold panel (TC1) by approximately +4.3°C on average and underestimated the hot panel (TC2) by approximately -2.5°C, producing a pooled RMSE of 4.11°C, with each per-target error falling within the range of other reported literature (Aragon et al., 2020; Han et al., 2020; Kelly et al., 2019). This gain error, missing from single-reference evaluations, is exactly the mechanism that a dual-target design is sensitive to (Lin et al., 2022) (Figure 2). The calibration reduced RMSE

to a range of 2.27–3.43°C across stable methods, a meaningful improvement over the factory baseline for crop-water index applications, though absolute errors remain non-trivial and warrant continued refinement. (Zhao et al., 2018; Das et al., 2021; Niu et al., 2020).

Temporal ELM performance provides partial support for Hypothesis 2. The static temporal variants (2A: 2.82°C, 2B: 2.60°C) demonstrated that models trained only on beginning and end of flight segments can reliably correct mid-flight temperatures. However, the interpolated variant (2C: 3.43°C) performed worse than the static variants, and the interpolated+FPA variant (2D: 5.39°C) performed worse than the factory baseline. This reversal likely reflects that for a long flight where sensor thermal state has largely stabilized, interpolating between begin and end coefficients introduces noise rather than capturing real drift. The bootstrap stability analysis confirmed that static temporal models are more reliable: the FPA-augmented static model (2B) had a narrow 95% stability interval of 2.57–2.78°C, indicating low sensitivity to which specific observations are in training (Figure 3).

Evidence for Hypothesis 4, concerning FPA-driven error, is mixed. Including FPA temperature as a predictor produced a consistent and statistically meaningful reduction in RMSE for the static temporal ELM (2A-2B: 2.82-2.60°C; bootstrap confirmed beneficial in 99.4% of resampled iterations, mean reduction 0.26°C). FPA inclusion also reduced RMSE for the generalized ELM (3A-3B: 2.51-2.27°C), though the Monte Carlo evidence is weaker (beneficial in 81% of iterations). In contrast, adding FPA to the interpolated temporal model (2C-2D) was harmful in 97.9% of bootstrap iterations, increasing RMSE from 3.43 to 5.39°C (Figure 1). Hypothesis 4 is therefore supported for static calibration models but not for interpolated ones; the benefit of FPA depends on the stability of the model structure to which it is added.

Hypothesis 3, regarding generalization across operating states, is supported. The generalized ELM trained on randomly sampled data spanning the full flight envelope achieved the best performance of any method: Method 3B (Generalized LWIR+FPA) at RMSE=2.27°C (44.8% reduction) and Method 3A (Generalized LWIR) at 2.51°C (38.9% reduction). Both values were lower than the best temporal static RMSE (2B: 2.60°C), indicating that training data spanning the full range of sensor and environmental states outperforms training restricted to stable flight-boundary periods. The Monte Carlo distributions showed stable performance across 200 random 80/20 partitions (95% CI: 1.53–3.10°C for 3B), confirming that this result is not dependent on a favorable random partition.

The variation of calibration error across wind speed and air temperature did not provide major context for understanding sensor behavior under field conditions (Table 2). With N=106 observations, the stratified analysis drew from approximately 22–48 observations per bin. The wind-speed pattern, low errors under high-wind conditions (RMSE=1.04°C) and high under medium-wind conditions (RMSE=2.96°C), is non-monotonic and does not follow a physically interpretable trend. Air temperature showed a similarly irregular pattern, with moderate air temperatures producing lower errors (RMSE=1.69°C) than either cool (2.77°C) or warm (2.23°C) conditions. With 22–48 observations per bin drawn from a single flight, these patterns are more likely driven by the confounding of environmental conditions with flight phase and sensor thermal state than by independent environmental effects. Based on the patterns I observed, no causal interpretation is warranted. Multi-flight datasets spanning a broader and more controlled range of meteorological conditions are required before environmental factors can be meaningfully evaluated as drivers of calibration error (Zhao et al., 2018; Kelly et al., 2019).

The downstream effects of calibration on canopy temperature and CWSI show that the RMSE reductions reported above translate into materially different agronomic signals of the same flight. The -7.55°C mean shift applied by Method 3B falls within the per-target error envelope of the factory baseline (TC1: +4.3°C; TC2: -2.5°C) and is consistent with correcting opposite-sign gain error across the canopy thermal distribution, a failure of uncalibrated UAV thermal imagery under changing flight conditions (Lin et al., 2022; Kelly et al., 2019). Because the pixel-wise correlation between raw and calibrated temperatures remains high ($r = 0.91$), the calibration preserves within-flight relative comparisons. The calibration is not creating a new structure but re-anchoring the existing structure to absolute temperature. The fitted Method 3B coefficients, $T_{cal} = 26.60 + 1.56 \cdot T_{raw} - 1.51 \cdot T_{FPA}$, decompose into two distinct corrections: the LWIR slope greater than one expands the compressed raw thermal range to match the radiometer reference (evident in the SD increase from 3.47°C to 5.26°C), and the negative FPA coefficient removes the bias as the focal plane warms during flight.

The Idso barley calculations provides an independent physical check (Idso, 1982; Kemanian et al., 2005). The raw mosaic places mean canopy pixel values more than 14°C above the well-watered theoretical reference and implausibly 8°C above the non-transpiring upper limit (Figure 6). The calibrated canopy distribution, by contrast, shows roughly half its mass between the well-watered lower limit and the non-transpiring upper limit, with most of the remainder above UL, a pattern consistent with severe late-season water stress in a heterogeneous dry-farmed stand under high VPD (Idso, 1982; Kemanian et al., 2005). These canopy-scale differences reinforce the operational consequence of factory-only calibration. The in-field target calibration corrects not only added noise, but systematic distortion that propagates into derived stress indices and would

impact evapotranspiration or irrigation decisions (Zhao et al., 2018; Niu et al., 2020; Das et al., 2021; Park et al., 2021).

From an operational standpoint, these findings imply that incorporating even a brief reference target imaging protocol at takeoff and landing can substantially reduce calibration error compared with relying on factory coefficients alone (Han et al., 2020; Han et al., 2021). For long flights where sensor thermal state stabilizes, the static temporal ELM with FPA (Method 2B) offers a robust, low-complexity correction with a narrow bootstrap stability interval. However, the best overall performance was achieved by the Generalized LWIR+FPA approach (Method 3B), which does not require segment assignment and benefits from the full range of training observations. When flight duration and conditions are representative of operational missions, I recommend method 3B as the primary calibration approach.

5. Limitations

The findings reported here derive from a single flight day under specific meteorological and operational conditions, which constrains the transferability of calibration coefficients to other dates, seasons, or geographic locations. Day-to-day variation in ambient temperature, solar loading, humidity, and wind regimes can shift sensor thermal dynamics and reference target behavior in ways that a single-day dataset cannot capture (Han et al., 2020; Messina & Modica, 2020). Consequently, while the temporal ELM and generalized ELM both demonstrated substantial error reduction relative to the factory baseline, I have not yet validated the extent to which these calibration relationships remain stable across flights conducted under different environmental envelopes. Because any residual slope miscalibration propagates into RMSE proportionally to the evaluation temperature range, absolute post-calibration RMSE values from this study are not directly comparable to those reported from single-panel or narrow-range

evaluations in the literature. I incorporate data from multiple flight days in future work, spanning a broader range of atmospheric and phenological conditions to assess calibration persistence and inform recommendations for recalibration frequency.

I based the current analysis on 106 synchronized radiometer-sensor pairs that I assigned to time-based begin (N=34), mid (N=26), and end (N=46) segments. While this is sufficient for robust model fitting and hypothesis testing, it does not yet support characterization of calibration behavior across diverse environmental conditions or flight dates. Lack of data under more variable conditions further prohibits the characterization of environmental factors (Table 2). The observed patterns suggest complex interactions that require multi-flight data to disentangle from random variation (Zhao et al., 2018; Kelly et al., 2019).

Photogrammetric blending artifacts caused by the low contrast of thermal imagery during feature matching add a residual uncertainty that the per-frame panel calibration cannot resolve (Ribeiro-Gomes et al., 2017). Future work should evaluate post mosaic correction to the per-frame application of the calibration prior to mosaicking, in order to quantify and reduce this portion of the error budget. Also in the mosaic application, I fit method 3B on panel temperatures of ~41–60 °C but applied to raw canopy pixels spanning roughly 33–47 °C (mean 39.97 °C, SD 3.47 °C), so most canopy pixels lie below the calibration training range. Over this ~5–8 °C extrapolation the radiance-temperature relationship should remain effectively linear. However, in future deployments I will add a cooler reference target that brackets canopy temperatures

Both the Apogee SI-121 infrared radiometers used as reference standards and the FLIR Boson microbolometer aboard the Altum-PT rely on manufacturer-provided calibration coefficients that I have not independently verified under controlled laboratory conditions for this study. While the Apogee radiometers are factory-traceable to NIST standards (Zhao et al., 2018), confirming their

performance across the specific temperature range and environmental conditions encountered during field operations would strengthen confidence in the reference measurements themselves (Aragon et al., 2020). Similarly, characterizing the Boson's response function and FPA-related drift behavior in a laboratory setting would help isolate sensor-intrinsic error sources from field-induced variability, clarifying why FPA parameterization provided only marginal improvement in this preliminary analysis. I plan independent laboratory testing of both instruments to substantiate manufacturer specifications and refine the error budget for subsequent calibration efforts.

6. References

Aragon et al. (2020). A calibration procedure for field and UAV-based uncooled thermal infrared instruments. *Sensors, 20*(11), 3316.

Das et al. (2021). UAV-thermal imaging: A technological breakthrough for monitoring and quantifying crop abiotic stress to help sustain productivity on sodic soils — A case review on wheat. *Remote Sensing Applications: Society and Environment, 23,* 100583.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*(6), 381–395.

Gerchow et al. (2025). Enhanced flight planning and calibration for UAV-based thermal imaging: Implications for canopy temperature and transpiration analysis. *Frontiers in Forests and Global Change, 8,* 1457762.

Han, X., et al. (2020). Field-based calibration of unmanned aerial vehicle thermal infrared imagery with temperature-controlled references. *Sensors, 20*(24), 7098.

Han, Y., et al. (2021). Calibration and image processing of aerial thermal image for UAV application in crop water stress estimation. *Journal of Sensors, 2021,* 5537795.

Idso, S. B. (1982). Non-water-stressed baselines: A key to measuring and interpreting plant water stress. *Agricultural Meteorology, 27*(1–2), 59–70.

Idso, S. B., Jackson, R. D., Pinter, P. J., Reginato, R. J., & Hatfield, J. L. (1981). Normalizing the stress-degree-day parameter for environmental variability. *Agricultural Meteorology, 24,* 45–55.

Kelly et al. (2019). Challenges and best practices for deriving temperature data from an uncalibrated UAV thermal infrared camera. *Remote Sensing, 11,* 567.

Kemanian, A. R., Stöckle, C. O., & Huggins, D. R. (2005). Transpiration-use efficiency of barley. *Agricultural and Forest Meteorology, 130*(1–2), 1–11.

Lin, D., Cui, X., Wang, Y., Yang, B., & Tian, P. (2022). Pixel-wise radiometric calibration approach for infrared focal plane arrays using multivariate polynomial correction. *Infrared Physics & Technology, 123,* 104110.

Messina, G., & Modica, G. (2020). Applications of UAV thermal imagery in precision agriculture: State of the art and future research outlook. *Remote Sensing, 12,* 1491.

Niu et al. (2020). Evapotranspiration estimation with small UAVs in precision agriculture. *Sensors, 20*(22), 6427.

Paciolla et al. (2025). Agronomic information extraction from UAV-based thermal photogrammetry using MATLAB. *Remote Sensing, 17*(15), 2746.

Park et al. (2021). Mapping very-high-resolution evapotranspiration from unmanned aerial vehicle (UAV) imagery. *ISPRS International Journal of Geo-Information, 10,* 211.

Ribeiro-Gomes, K., Hernández-López, D., Ortega, J. F., Ballesteros, R., Poblete, T., & Moreno, M. A. (2017). Uncooled thermal camera calibration and optimization of the photogrammetry process for UAV applications in agriculture. *Sensors, 17*(10), 2173.

Sobrino, J. A., Jiménez-Muñoz, J. C., Sòria, G., Romaguera, M., Guanter, L., Moreno, J., Plaza, A., & Martínez, P. (2008). Land surface emissivity retrieval from different VNIR and TIR sensors. *IEEE Transactions on Geoscience and Remote Sensing, 46*(2), 316–327.

Williams, J. L., Sherman, J. D., Lamb, P., Cook, J., Lachowiec, J. A., & Bourgault, M. (2022). Relationships between roots, the stay-green phenotype, and agronomic performance in barley and wheat grown in semi-arid conditions. *The Plant Phenome Journal, 5*(1), e220050.

Zhao, T., Niu, H., Anderson, A., Chen, Y., & Viers, J. (2018). A detailed study on accuracy of uncooled thermal cameras by exploring the data collection workflow. *Proceedings of SPIE, 10664,* 106640F.