Energy & Buildings 369 (2026) 117983

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild

Adapting ISO 9869-1 for post-retrofit low U-value walls: Evidence from 130 days of continuous in-situ monitoring

Hu Du*, Alexander Pilkington

School of Engineering and Built Environment, Liverpool John Moores University, Liverpool, United Kingdom

ARTICLE INFO

Keywords:
In-situ U-value
ISO 9869-1
Heat flow meter
Test duration
Wind-driven rain
External wall insulation

ABSTRACT

Large scale retrofit programmes require a practical means of verifying that newly insulated walls perform as designed. ISO 9869-1, the international heat flow meter method, permits test durations as short as 72 hours, a threshold established when uninsulated walls lost heat far more rapidly than today's high performance retrofits. This study evaluates the suitability of the standard for modern post retrofit construction and proposes evidence based adaptations.

A recently insulated cavity brick wall in Liverpool, UK (design U-value 0.18 W/(m^2 K)), was monitored at two adjacent points over 130 days. Heat flux and indoor and outdoor temperatures were logged at 10-minute intervals alongside on site weather data. The Average Method was applied to all 13,687 candidate time windows to examine the influence of test duration, starting hour, rainfall, and other environmental factors.

The 72-hour minimum performed poorly: only 5–10% of windows satisfied the convergence criteria, and among these the mean absolute error against long duration references was approximately 32%. A minimum of 14 days was required for indicative results, and 28 days for accuracy within ±5%. Starting hour was statistically irrelevant once windows exceeded three days. Moisture proved decisive: wet periods inflated measured U-values by 13–16%, with recovery of 4–6 mW/(m^2 K) per consecutive dry day. Measured values of 0.206 and 0.188 W/(m^2 K) exceeded the design value by only 0.026 and 0.008 W/(m^2 K) in absolute terms.

Three adaptations to ISO 9869-1 are proposed: a wall-type-dependent minimum duration, mandatory dry-spell pre-conditioning, and a combined absolute and relative tolerance for design measurement agreement. Together these provide a measurement based quality assurance route for large scale retrofit delivery.

1. Introduction

Buildings account for approximately 40% of final energy use and a comparable share of CO_2 emissions in Europe, and opaque-envelope performance is a principal determinant of heating demand [1,2]. UK policy now turns this physics challenge into a delivery one. The Warm Homes Plan commits £15 billion of public investment to upgrade up to 5 million homes by 2030, with consumer protection central to delivery across homeowners, landlords, local authorities and housing providers [3]. That emphasis has been reinforced by the 2025 response to poor-quality solid-wall insulation installed under inherited Energy Company Obligation 4 (ECO4) and Great British Insulation Scheme (GBIS) routes, in which installers were suspended, Ofgem (the Office of Gas and Electricity Markets, the independent UK energy regulator) began writing to affected households, and responsible installers were required to remedy faulty work at no cost to consumers [4]. Subsequent audits reported major technical non-compliance in 92% of sampled external-wall-insulation installations and 27% of sampled internal-wall-insulation installations under ECO4/GBIS [5]. In this context, reliable post-retrofit U-value verification is no longer a secondary monitoring exercise; it is evidence of whether retrofit investment has delivered the heat-loss reductions and effective insulation it was funded to provide.

The building-physics literature explains why this verification cannot rely on design calculations alone. Discrepancies between designed and measured thermal transmittance, commonly termed the "performance gap," have been repeatedly documented for both uninsulated and retrofitted envelopes [6–9]. In a meta-analysis of past in-situ studies on envelopes with design U-values below 0.3 W/(m^2 K), O'Hegarty et al. [8] reported that 93% of tests showed underperformance; in their own campaign of 10 in-situ measurements on highly insulated envelopes (design U-value < 0.2 W/(m^2 K)), 9 of 10 underperformed, with an average deviation exceeding 100%. However, percentage deviations become amplified as the denominator becomes small, and so

* Corresponding author.
E-mail address: H.Du@ljmu.ac.uk (H. Du).

https://doi.org/10.1016/j.enbuild.2026.117983
Received 1 May 2026; Received in revised form 29 June 2026; Accepted 20 July 2026
Available online 24 July 2026

Nomenclature

Latin symbols

D	Test window duration (h or days)
D_{ref}	Duration of the longest converged window (days)
$I_A(\Theta)$	Orientation-specific airfield wind-driven rain index (mm)
$I_{A,\mathrm{ann}}(\Theta)$	Annualised airfield wind-driven rain index (mm yr^{-1})
I_k	Permutation importance of feature k (–)
k_{dry}	Drying-rate coefficient; slope of U vs dry-spell length (W m^{-2} K^{-1} day^{-1})
M	Number of permutation repeats in Random-Forest importance (–)
n	Number of time-steps in a candidate analysis window (–)
N_{days}	Duration of the weather data subset used for annualisation (days)
q	Heat flux density measured at the internal wall surface (W m^{-2})
R_D	Cumulative surface thermal resistance over window of duration D (m^2 K W^{-1})
r_h	Hourly rainfall total (mm)
T_{in}	Indoor air temperature (°C)
T_{out}	Outdoor air temperature (°C)
T_{si}	Indoor wall surface temperature (°C)
T_{se}	Outdoor wall surface temperature (°C)
U	Thermal transmittance (U-value) computed by the Average Method (W m^{-2} K^{-1})
U_0	Dry-baseline intercept in the drying regression (W m^{-2} K^{-1})
U_{design}	Design U-value per BS EN ISO 6946 (W m^{-2} K^{-1})
U_{ref}	Long-duration reference U-value; median of the longest converged windows (W m^{-2} K^{-1})
V_h	Hourly mean wind speed (m s^{-1})

Greek symbols

ΔT	Temperature difference $T_{\mathrm{in}} - T_{\mathrm{out}}$ (°C)
ΔU_{abs}	Absolute U-value tolerance in the dual-tolerance criterion (W m^{-2} K^{-1})
δ_{rel}	Relative U-value tolerance in the dual-tolerance criterion (–)
ε	Relative error of a candidate window vs the reference U-value (–)
ϕ_h	Hourly mean wind direction (degrees; north = 0°)
τ_{dry}	Dry-spell length: consecutive dry days before the window centre (days)
Θ	Wall orientation angle (degrees; north = 0°)

Abbreviations

DRI	Driving Rain Index
EWI	External Wall Insulation
HFM	Heat Flow Meter
MAE	Mean Absolute Error
RH	Relative Humidity
WDR	Wind-Driven Rain

low U-value fabrics should be assessed using both absolute and relative differences. These findings motivate the routine use of in-situ measurement to verify as-built performance, certify retrofits and calibrate energy models.

The reference method is the heat flow meter (HFM) technique standardised in BS ISO 9869-1:2014 [10], wherein heat flux and indoor/outdoor temperatures are logged continuously until three convergence criteria are satisfied: a minimum duration of 72 hours, a 24-hour stability criterion and a first/last two-thirds comparison (both with a 5% tolerance). Although internationally adopted, the method has persistent practical limitations that have received detailed treatment in the literature.

Duration. Multiple studies report that the nominal 72-hour minimum is often insufficient to satisfy the ISO 9869-1 convergence criteria for many wall types and field boundary conditions, particularly for insulated or heavyweight constructions [11–13]. Heavyweight and highly insulated walls may require 7–21 days [8,14–16], and the third convergence criterion in particular is acknowledged to be excessively restrictive [16,17]. Gaspar et al. [18] showed that façades with U-values below 0.5 W/(m^2 K) require substantially longer test durations to avoid systematic deviations, and a subsequent review of HFM minimum-test-duration criteria [19] found no consensus on a wall-type-dependent floor. For high-performance retrofitted walls the heat-flux signal is small and the signal-to-noise ratio compounds the problem: Farmer et al. [20] applied the Average Method at successive stages of a full fabric retrofit on a solid-wall Victorian dwelling situated within the Salford Energy House environmental chamber, and reported that in-situ U-values (down to 0.22 W/(m^2 K) at the post-retrofit stage) can be obtained with shorter durations where steady-state boundary conditions and large indoor outdoor temperature difference were imposed in the chamber.

Starting conditions. The standard notes that the 72-hour minimum applies only “if the temperature is stable around the HFM” [10], and recommends avoiding direct solar exposure. Guidance on the *time of day* at which measurements should begin is, however, implicit rather than prescriptive. Practitioners routinely start tests at night to limit solar transients, but whether the start hour materially affects the time-averaged result of a multi-day test does not appear to have been systematically tested on a single wall.

Moisture. Masonry is known to exhibit transient thermal properties under wetting–drying cycles. Laboratory and field work on historic masonry [14,15,21] and on envelopes exposed to alternating heat flux [22,23] have shown that moisture content elevates the apparent U-value. Juricic et al. [24] also demonstrated that natural weather variability introduces significant uncertainty into in-situ characterisation, yet quantifications using wind-driven rain (WDR) as the independent driver remain sparse, particularly for maritime-temperate climates in which the ISO 9869-1 assumption of quasi-steady conditions is routinely violated. This is a particular concern for retrofitted walls in exposed UK locations, where WDR loading on newly clad surfaces may differ appreciably from the pre-retrofit condition.

Table 1 summarises the measurement parameters of published in-situ HFM studies on external walls, ranked by monitoring duration from longest to shortest. The table indicates the wide variation in study length (from 72 h to 365 days), wall typology and U-value range, and places the present 130-day study, which is one of the longest published continuous monitoring for a UK wall, within the broader evidence base.

1.1. Aim and structure of this study

The aim of this research is to determine how an ISO 9869-1 Average-Method test should be planned and interpreted for a post-retrofit low U-value wall. Four linked analyses structure the study, moving from the duration needed to obtain a stable signal to the environmental and experimental factors that explain the remaining variation:

1. **Length of monitoring**: how the probability of ISO 9869-1 convergence, and the U-value error relative to a long-duration reference, scale with window length from 72 h to the longest complete windows available.
2. **Starting time**: whether the *time of day* at which an Average-Method test begins measurably affects the resulting U-value.
3. **Moisture impact**: how much moisture ingress, characterised by WDR and dry-spell length, biases the measured U-value.

Table 1
Summary of published in-situ HFM studies on external walls: measurement duration, equipment, wall type, location and U-value range. Studies are ranked by monitoring duration (longest first); the present study is shown in bold.

Reference	Country	Wall type	U-value (W/m^2 K)	Duration	Equipment
Lee et al. [25]	Korea	Insulated exterior	∼0.15	365 days[a]	HFP01, Omega
Present study	**UK**	**Cavity brick + EWI**	**0.19–0.21**	**130 days**	**greenTEG wireless**
Ghazi Wakili et al. [26]	Switzerland	Retrofitted brick/stone	0.382	43–79 days	HFM
Flores Larsen et al. [22]	Argentina	Cavity brick masonry	0.29–1.73	23–57 days	HFP01, HOBO U12-014
Rodler et al. [27]	France	Timber/historic	0.3–0.5	33 days	Thermocouple
Gori & Elwell [28]	UK	Solid brick (all seasons)	0.59–1.92	3–30 days	HFP01, Campbell
Atsonios et al. [12]	Greece	Insulated wall	0.33–1.25	5–28 days	TRSYS01, Campbell
Choi & Ko [17]	Korea	Exterior walls	0.17–0.48	21 days	Datalogger
Gaspar et al. [18]	Spain	Low U-value façades	0.10–0.23	72–504 h	HFP01, gSKIN
Evangelisti et al. [29]	Italy	Various orientations	0.24–0.37	18 days	Mixed
Rasooli & Itard [30]	Netherlands	Various walls	0.27–5.31	14–16 days	HFP01
Evangelisti et al. [13]	Italy	Various	∼1.31	3–15 days	Mixed
Juricic et al. [24]	France	Various	1.3–1.6	2–15 days	Modelling
Biddulph et al. [31]	UK	Solid brick (93 walls)	1.16	14 days	Hukseflux HFP01
Lucchi [14]	Italy	Historic brick masonry	0.94–2.10	14 days	Ahlborn ALMEMO
Ahmad et al. [32]	Saudi Arabia	Hollow precast concrete	1.459–1.803	6–14 days	HFP01, Campbell
Kim et al. [33]	Korea	Insulated curtain wall	0.52–1.06	72 h–14 days	T-type TC
Genova & Fatta [15]	Italy	Historic calcarenite	0.56–1.07	72 h–11 days	Ahlborn ALMEMO
Marshall et al. [6]	UK	Brickwall	0.462–3.835	7–9 days	HFP01
Gaspar et al. [19]	Spain	Various (review)	0.23–1.25	60–192 h	HFP01, Campbell
Bienvenido-Huertas et al. [34]	Spain	140 wall typologies	0.09–0.32	7 days	Mixed
O'Hegarty et al. [8]	Ireland	Highly insulated	0.2–0.3	7 days	TRSYS01
Li et al. [7]	UK	Solid brick	1.29–1.34[b]	3–7 days	HFP01
Evangelisti et al. [35]	Italy	Various wall types	0.54–1.58	4–6 days	Mixed
Ficco et al. [36]	Italy	Various existing	1.35–1.68	72–120 h	Mixed
Gaspar et al. [11]	Spain	Various façades	0.09–2.35	72 h	HFP01, Campbell
Farmer et al. [20][d]	UK	Solid wall + retrofit	0.22–0.33[c]	72 h	HFP01
Bergero [37]	Italy	Cavity brick/masonry	0.73–2.54	72 h	HFM

[a] Year-long monitoring; analysis windows of 3–15 days.
[b] Mean of brick and stone wall samples.
[c] Post-retrofit values.
[d] Conducted at the Salford Energy House test facility under controlled steady-state conditions, not field in-situ. TC = thermocouple.

4. **Significance of features**: which experimental and environmental factors, in rank order, control the scatter in 14-day U-values on this wall. The 14-day base window is adopted from the length-of-monitoring results in Section 3, which identify it as the shortest duration at which the U-value distribution becomes statistically stable on this wall (mean absolute error 10–12%, interquartile range ±0.02 W/(m^2 K)) and which also matches the moisture-recovery time scale used in the dry-spell analysis.

The study draws on 130 days of continuous 10-minute heat-flux and temperature data from two closely spaced sensor stations on the same east-facing post-retrofit wall, combined with on-site 5-minute weather data. The 130-day record makes this one of the longest published continuous in-situ monitoring records for a UK retrofitted wall, and, to the authors' knowledge, among the first to examine monitoring length, starting time, WDR, moisture and design agreement together on the same high-performance envelope. The analysis proceeds from measurement duration to moisture context, and then to starting time and feature significance. This order is deliberate: duration establishes whether a stable window exists at all, moisture explains why otherwise similar windows can differ, starting time tests a common scheduling assumption once multi-day averaging is in place, and feature significance identifies the measured conditions that control the remaining scatter. The same framework is also used to interpret the design–measurement difference between the measured U-value (0.19–0.21 W/(m^2 K)) and the BS EN ISO 6946 design value (0.18 W/(m^2 K)), where a relative difference of up to 14% corresponds to a marginal absolute difference of approximately 0.02 W/(m^2 K) for this low U-value wall.

The remainder of the paper is organised as follows. Section 2 describes the measurement setup and analysis pipeline. Section 3 presents the four linked analyses: length of monitoring, moisture impact, starting time and significance of features. Section 4 discusses the implications for ISO 9869-1 practice and compares the findings with the literature. Section 5 concludes.

2. Materials and methods

2.1. Methodology overview

Fig. 1 summarises the study. The work proceeds in five steps that are aligned with BS ISO 9869-1:2014 [10]: (i) defining the protocol question from the literature; (ii) selecting a representative post-retrofit wall and instrumenting it with two heat-flux stations; (iii) continuous monitoring of heat flux, surface and air temperatures, alongside on-site weather; (iv) systematic analysis of the four controlling factors (duration, starting time, moisture and feature significance); and (v) cross-validation between the two stations. The same 130-day dataset feeds every analysis, so the results can be compared on a like-for-like basis. The analysis pipeline (the Average-Method estimator, the convergence test and the four post-processing analyses described in Section 2.5) is the methodological core of the study, while the wall, site and instrumentation described next define the boundary conditions under which that pipeline is exercised. The remainder of this section describes the wall and site, the instrumentation, the validation of the data pipeline, and the analysis used for each of the four questions.

2.2. Wall and site

The test wall is an east-facing external wall of a post-2010 dwelling in Liverpool, UK (53.41°N, 2.98°W). It was originally a cavity-brick envelope and was later upgraded with full-fill cavity insulation, external wall insulation (EWI) and brick-slip cladding, giving a design U-value of 0.18 W/(m^2 K) per BS EN ISO 6946 [38]. This combination of high-performance, thermally massive and multi-layered build-up is increasingly common in UK retrofit programmes, and it is a demanding

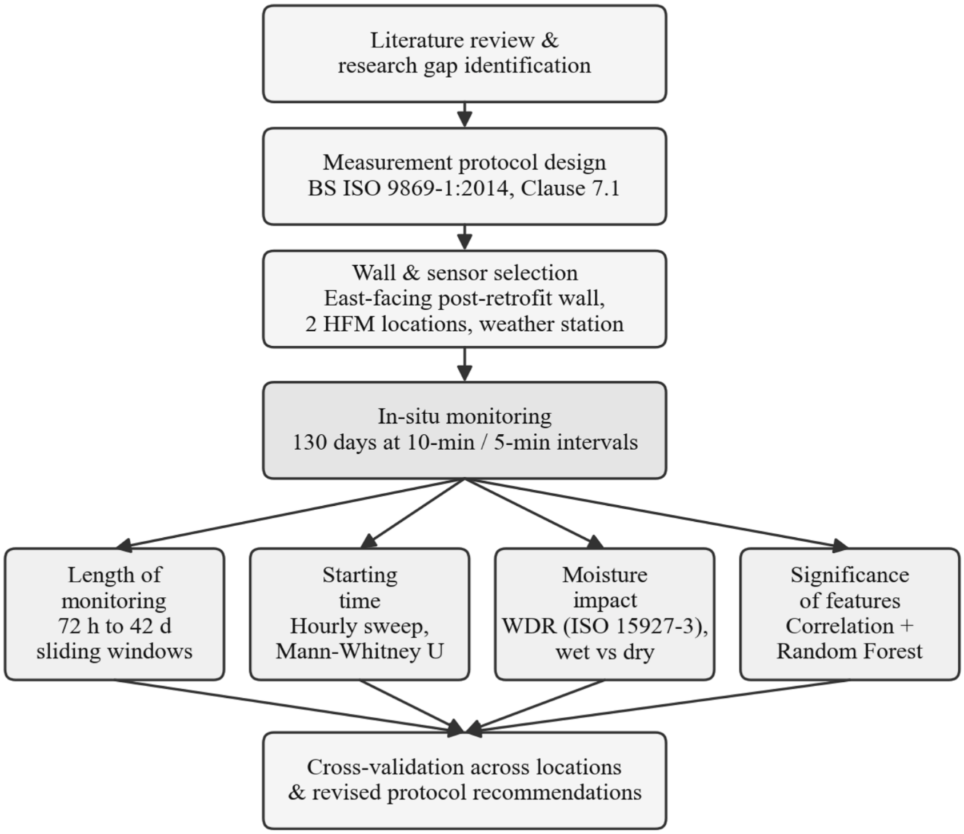

Fig. 1. Study workflow from measurement framework to cross-validation across two closely spaced sensor stations on the same east-facing post-retrofit wall. Length of monitoring, starting time, moisture impact and significance of features are evaluated on the same 130-day dataset and replicated at the second sensor station.

case for the Average Method [8,18,20]: the small heat-flux signal lowers the signal-to-noise ratio, the high thermal mass lengthens the wall's response time, and the layered construction introduces hygrothermal complexity. The wall is therefore representative of the post-retrofit UK dwellings for which in-situ U-value verification is now being requested.

The monitoring period ran from 2 December 2025 to 10 April 2026, a period of 130 calendar days that covers heating-dominated winter and early-spring conditions. This is substantially longer than most published single-wall studies, which typically span 7–30 days [1,39], and so allows seasonal convergence behaviour to be examined systematically. Liverpool has a maritime-temperate climate (Köppen Cfb) with substantial wind-driven rain (WDR) exposure, providing the natural moisture forcing required for the moisture-impact analysis.

2.3. Instrumentation

Two adjacent heat-flux stations were installed on the same east-facing wall at the same floor level, providing within-wall replication under shared boundary conditions. The two stations were deliberately placed close together, rather than far apart, because the objective of this study is not to map facade-scale spatial heterogeneity but to quantify the *repeatability* of an ISO 9869-1 Average-Method test under nominally identical conditions: two sensors exposed to the same indoor and outdoor environment provide a controlled estimate of the measurement-to-measurement variability that a practitioner would encounter, isolated from differences in orientation, shading or wind exposure. The two stations were both located on the clear central field of the wall, away from the reveals, lintel, floor junction and any EWI fixing positions; the distance from each plate to the nearest corner is several times the total wall thickness, so the measurement points satisfy the one-dimensional, edge-effect-free heat-flow assumption underlying Eq. (1). The close agreement between the two independently instrumented stations throughout the campaign (Fig. 6 and the rolling time series in Section 3) is itself evidence that the heat flow at both points is one-dimensional and free of localised thermal bridging:

- **Location 1**: heat-flux plate plus indoor surface and air temperature sensor (SM-10284), paired with an outdoor surface and air temperature sensor (SM-10292).
- **Location 2**: equivalent indoor pair (SM-10290) and outdoor pair (SM-10289).

All channels were logged at 10-minute resolution, giving 144 samples per channel per day and 18,549 time-stamped records in the merged U-value dataset. Heat-flux plates were bonded to the internal surface with a thin layer of high-thermal-conductivity paste, and sensor positions were chosen to avoid thermal bridges, surface irregularities and direct solar exposure, in line with ISO 9869-1 guidance [10,40]. Fig. 2 shows the sensor mounting and a section view of the wall; Fig. 3 shows the building east elevation with the outdoor sensor positions and the on-site weather station; and Fig. 4 shows the indoor sensors and logging gateway.

An on-site weather station logged wind speed and direction, rainfall, air temperature, relative humidity, solar radiation and surface wetness at 5-minute resolution (37,728 records). Sensor GPS coordinates were recorded, and nearby building footprints were retrieved from OpenStreetMap to characterise the micro-sheltering context. Table 2 lists every channel, its sensor, accuracy and logging interval; this table contains the information needed to replicate the measurement chain.

2.4. Sensor validation

Before deployment, the in-house Python implementation of the Average Method (Eq. (1)) was checked against the manufacturer's

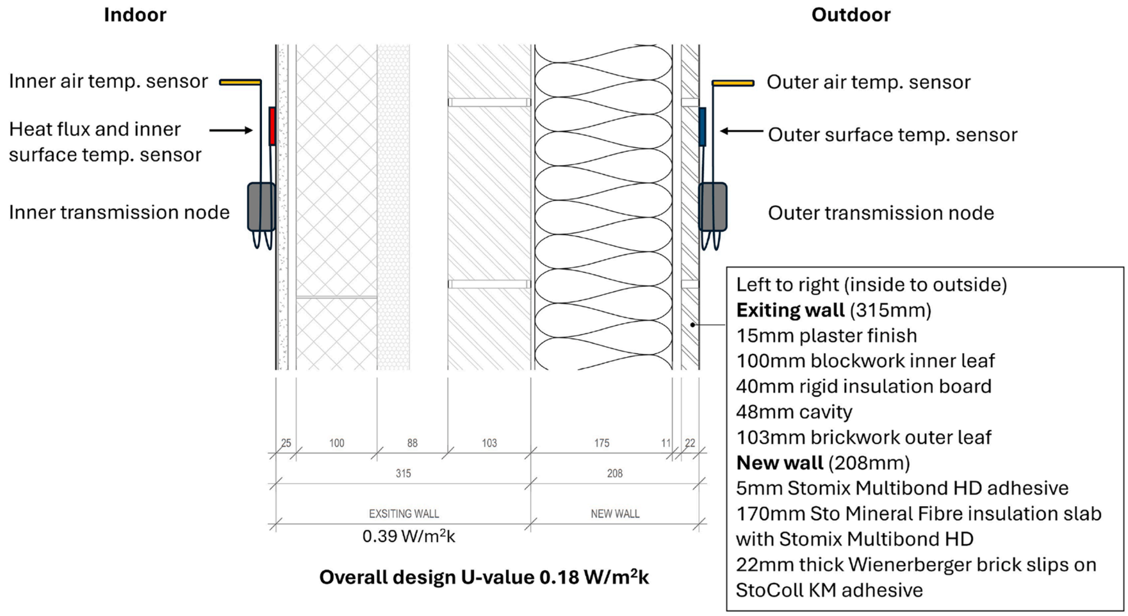

Fig. 2. Sensor installation schematic including section view of the post-retrofit external wall (cavity brick + insulation + EWI + brick slip), showing indoor/outdoor surface and air temperature probes bonded to both sides of the wall and the heat-flux plate bonded to the internal surface.

Fig. 3. Outdoor nodes mounted on the building east façade. A. HOBO RX3000 weather station Advanced Kit with BF5 diffuse radiation sensor (71 meters away from the site, on the roof of nearby building). B. HOBOnet Wireless Ultrasonic Wind Speed & Direction Sensor for local wind measurement. C. greenTEG outer surface temperature sensor. D. greenTEG outer air temperature sensor.

reference software (greenTEG) on an identical 72-hour controlled-laboratory dataset. The two pipelines agreed to better than 0.3% in the computed U-value and 0.5% in the 24-hour moving mean of heat flux (Fig. 5), so any differences reported in Section 3 are attributable to physical and protocol factors rather than to software artefacts.

A combined standard uncertainty for the Average-Method U-value was then estimated by propagating the manufacturer accuracies in Table 2 through Eq. (1), following the ISO/IEC Guide 98-3 (GUM) law of propagation of uncertainty [41]; the full propagation equations are given in Appendix A. Under typical winter conditions on this wall ($q \approx 2$–3 W/m^2, $\Delta T \approx 10$ °C), the dominant terms are heat-flux plate accuracy (± 0.09 W/m^2, contributing ∼3–4%) and the air-temperature pair (each ± 0.1 °C, together contributing ∼1.4% via ΔT). The combined relative standard uncertainty in U is therefore about 4–5% for a single 14-day window — in practical terms only about ± 0.01 W/(m^2 K) on a 0.2 W/(m^2 K) wall — and reduces with $1/\sqrt{n}$ for longer integrations. This single-window combined uncertainty is carried through the Results as an instrumentation-error band: any measured effect substantially larger than this band cannot be attributed to sensor accuracy alone, and the explicit comparisons for the moisture and design–measurement effects are made where those results are presented (Sections 3.3 and 4).

2.5. Analysis

All analysis was performed in Python 3.14 using `numpy`, `pandas`, `scipy`, `scikit-learn` and `matplotlib`. Scripts and data will be made available alongside the dataset in the LJMU Data Repository (see Data Availability), so the pipeline described below can be reproduced end-to-end.

Fig. 4. Indoor nodes. A. Heat-flux plate. B. Air temperature probes. C. Wireless communication node. D. Data logging gateway.

Table 2
Instrumentation summary. U-value channels were logged at 10-minute intervals; weather station channels at 5-minute intervals. BF5 accuracy figures are at 95% confidence (spectral range 400–700 nm).

Variable	Unit	Accuracy	Interval	Standard
greenTEG UVAL Wireless System (gOMS II)				
Heat flux (q)	W/m^2	±0.09 W/m^2	10 min	ISO 9869-1, ASTM C1046
Surface temperature	°C	±0.1 °C	10 min	ISO 9869-1, ASTM C1046
Air temperature	°C	±0.1 °C	10 min	ISO 9869-1, ASTM C1046
HOBOnet Ultrasonic Wind Sensor (RXW-WCG)				
Wind speed	m/s	±0.8 m/s or ±4%	5 min	BS EN ISO 15927-3
Wind direction	°	±2° (>3 m/s)	5 min	BS EN ISO 15927-3
Onset 0.2 mm Rain Gauge (S-RGB-M002)				
Rainfall	mm	±1.0% (≤20 mm/h)	5 min	BS EN ISO 15927-3
Delta-T BF5 Sunshine Sensor				
Total solar radiation	W/m^2	±5 W/m^2 or ±12%	5 min	—
Diffuse solar radiation	W/m^2	±20 W/m^2 or ±15%	5 min	—
Sunshine status	—	±10% (WMO threshold)	5 min	WMO
HOBO RX3000 Weather Station smart sensors				
Air temperature (ext.)	°C	±0.20 °C	5 min	WMO-No. 8
Relative humidity	%	±2.5%	5 min	WMO-No. 8
Barometric pressure	mbar	±3.0 mbar	5 min	WMO-No. 8
Surface wetness	%	±5%	5 min	—

2.5.1. *U-value estimation (average method)*

This paper focuses on the Average Method because it is the reference protocol specified by ISO 9869-1 Clause 7, requires only standard heat-flux and temperature inputs without any parameter fitting, and is the method routinely used by practitioners and accredited testing bodies for retrofit verification. The central question of this study is therefore precisely how this widely deployed protocol behaves on a low U-value retrofit wall. Dynamic, Bayesian and dual-heat-flux-meter alternatives [27,28,31] applied to the same dataset will be reported in a separate paper.

For each candidate window $[t_0, t_0 + D]$ the Average-Method U-value is computed per ISO 9869-1 Clause 7.1 as

$$U = \frac{\sum_{j=1}^{n} q_j}{\sum_{j=1}^{n} \left(T_{\mathrm{in},j} - T_{\mathrm{out},j}\right)}, \tag{1}$$

with q_j the heat flux (W/m^2) and $T_{\mathrm{in},j}$, $T_{\mathrm{out},j}$ indoor and outdoor temperatures (°C). Convergence was assessed using the cumulative surface resistance $R_D = \sum(T_{\mathrm{si}} - T_{\mathrm{se}})/\sum q$ over each window, where T_{si} and T_{se} are the indoor and outdoor surface temperatures, as implemented in ISO 9869-1. A window is declared *converged* if all three Clause 7.1 criteria are simultaneously satisfied:

$$\text{(C1) Minimum duration:} \quad D \geq 72\ \mathrm{h}, \tag{2}$$

$$\text{(C2) 24-hour stability:} \quad \frac{\left|R_D - R_{D-24\mathrm{h}}\right|}{R_D} \leq 5\%, \tag{3}$$

$$\text{(C3) First/last two-thirds test:} \quad \frac{\left|R_{\mathrm{first}\,2D/3} - R_{\mathrm{last}\,2D/3}\right|}{R_D} \leq 5\%. \tag{4}$$

The relative error of any candidate window against the reference is

$$\varepsilon = \frac{U - U_{\mathrm{ref}}}{U_{\mathrm{ref}}}. \tag{5}$$

Reference U-values were taken as the median U-value of the longest converged windows at each location,

$$D_{\mathrm{ref}} = \max\{D : (\mathrm{C1}) \wedge (\mathrm{C2}) \wedge (\mathrm{C3})\},$$
$$U_{\mathrm{ref}} = \mathrm{median}_{t_0}\{U(t_0, D_{\mathrm{ref}}) : (\mathrm{C1}) \wedge (\mathrm{C2}) \wedge (\mathrm{C3})\}, \tag{6}$$

yielding $U_{\mathrm{ref,Loc1}} = 0.206$ W/(m^2 K) and $U_{\mathrm{ref,Loc2}} = 0.188$ W/(m^2 K). The measured long-duration reference U_{ref} of Eq. (6), rather than the BS EN ISO 6946 design value, is used as the benchmark for the

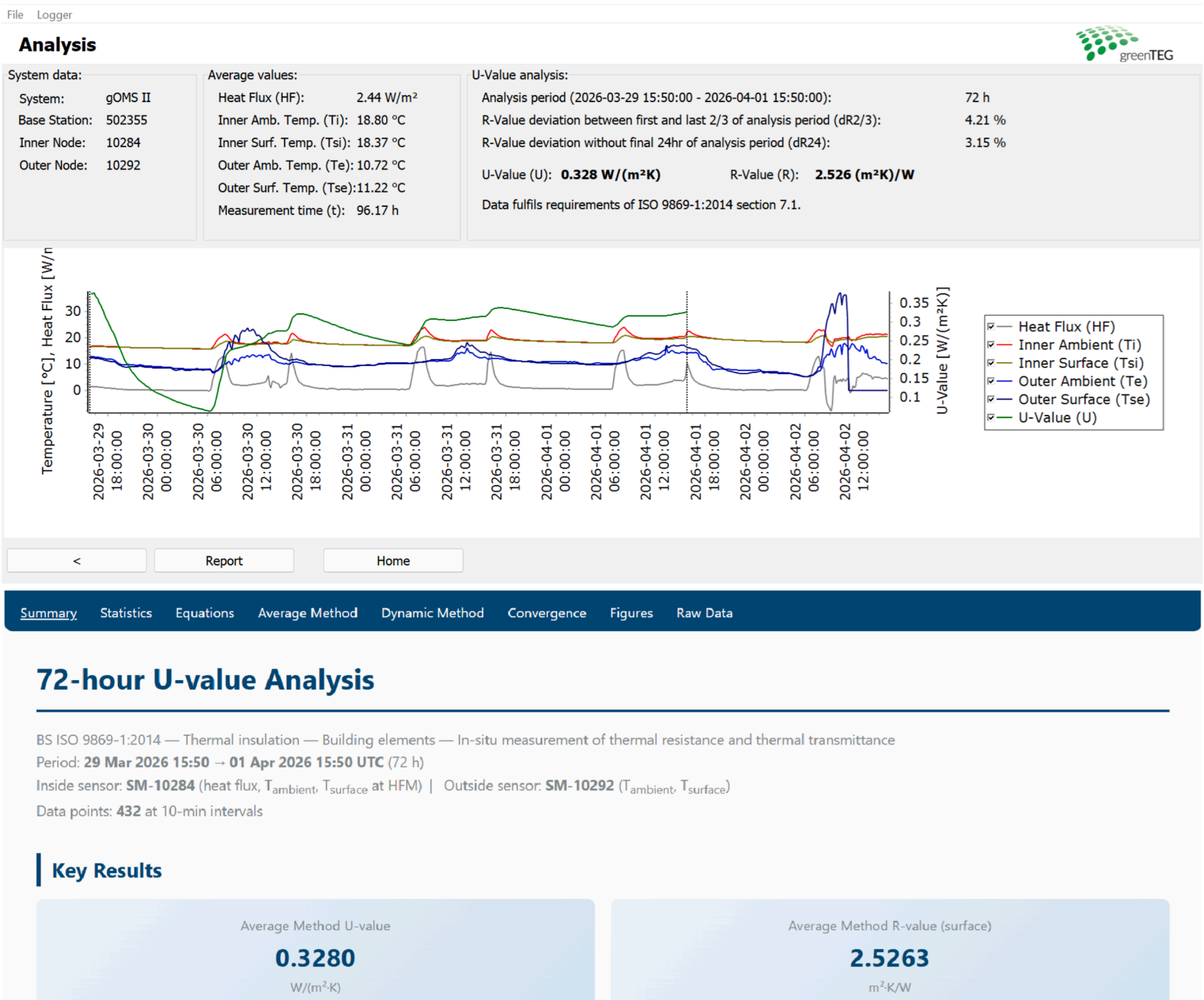

Fig. 5. Cross-validation of the in-house Python implementation (bottom) of the ISO 9869-1 Average Method against the greenTEG manufacturer software (top) on an identical 72-hour controlled dataset. Green line shows the moving-mean U-value.

duration, starting-time and moisture analyses, and this choice is deliberate. A real wall is not a single homogeneous slab: the as-built section combines brick, mortar, cavity fill, EWI, brick-slip cladding and discrete fixings, together with mortar joints and movement or expansion joints, and the design value, computed from nominal layer properties under a one-dimensional steady-state assumption, cannot represent this composite build-up or the inevitable gap between design and installation. The long-duration measured reference, by contrast, is the value that the same instrument returns on the same point of the same wall once the transients have averaged out, and it is therefore the correct internal benchmark for questions about the measurement process itself. The same reasoning is why the two heat-flux stations were placed close together on the clear field of the wall: keeping the boundary conditions and the composite cross-section as nearly identical as possible between the two points isolates measurement repeatability from facade-scale heterogeneity. The independent design value of 0.18 W/(m^2 K) is reserved for the separate *performance-gap* question in Section 4, where the design-to-measurement difference is the quantity of interest, consistent with the convention of O'Hegarty et al. [8]. The two location references differ by ~0.018 W/(m^2 K) (~10%), which is reported in Section 4 as an indicative measure of within-wall variability between two points on the same facade rather than treated as a discrepancy to be reconciled. To check that the relative-error analysis is not circular (the reference is computed from the same dataset that is being judged), U_{ref} was recomputed on the first half and the second half of the monitoring period separately. The two halves agreed to within ±0.005 W/(m^2 K) (~2.5%), indicating that the rank ordering of duration, starting-time and moisture effects is robust to the choice of reference partition.

2.5.2. *Length of monitoring*

The length-of-monitoring analysis tests the central protocol question: whether the 72-hour minimum is long enough to average out thermal mass, sensor noise and environmental transients on this kind of wall. Window durations D were therefore swept in 24-hour steps from 72 h to the maximum complete window available at each location. The main figures focus on the practice-relevant range of 72 h to 42 days; longer windows were retained to define the reference U-values and the long-duration error thresholds. For every (D, t_0) pair we recorded the U-value (Eq. (1)), whether the window was converged under Clause 7.1 (Eqs. (2)–(4)), and the relative error ε against the long-duration reference (Eq. (5)). Mean, 5th/95th percentiles of $|\varepsilon|$ and convergence rate were tabulated per duration and per location.

The number of candidate windows decreases approximately linearly with D, from ~229 at $D = 3$ days to ~211 at 14 days, ~183 at 28 days, ~99–109 at 65–70 days, and a single window at 121 days (both locations combined; 130-day study period, 1-day stride, after rejecting any window with $|\overline{\Delta T}| < 3\,^\circ$C). Summed across all durations $D = 3$–121 days and both locations, the sweep produces 13,687 candidate windows. Reported error statistics for the longest durations therefore have wider sample uncertainty and should be read as indicative.

2.5.3. Starting time

The starting-time analysis tests whether the common preference for night-started measurements still matters once several diurnal cycles have been averaged. For each duration $D \in \{3, 5, 7, 10, 14, 21\}$ days, every possible start hour $h \in \{0, \ldots, 23\}$ was evaluated. Converged windows were grouped into *night starts* (18:00–06:00) and *day starts* (06:00–18:00), and the two groups were compared with two-sided Mann–Whitney U tests [42]. A Kruskal–Wallis test [43] across all 24 hours acted as an omnibus check against any single hour-of-day differing from the rest. Non-parametric tests were used because converged U-values at short durations are skewed and not reliably normal. Effect sizes were reported as Cohen's d [44] on the standardised differences, where $|d| < 0.2$ is conventionally negligible, 0.2–0.5 small, 0.5–0.8 medium and >0.8 large; this lets a statistically detectable effect be distinguished from a practically meaningful one. The family of Mann–Whitney p-values across the six durations was also checked under a Benjamini–Hochberg correction [45] at the 5% false-discovery rate; no test became significant after correction, consistent with the uncorrected $p > 0.7$ values.

2.5.4. Moisture impact and WDR

Wind-driven rain on the east-facing wall was computed following BS EN ISO 15927-3:2009 [46]. Total rainfall on its own does not indicate whether the rain was actually driven onto the instrumented surface, whereas the wall's measured response depends on orientation-specific wetting. Five-minute rainfall was therefore summed to hourly totals, wind speed averaged hourly, and wind direction circularly averaged. The orientation-specific airfield index for a vertical surface facing direction Θ was then computed as

$$I_A(\Theta) = \tfrac{2}{9} \sum_h V_h \, r_h^{8/9} \, \max\{\cos(\phi_h - \Theta), 0\}, \tag{7}$$

where V_h is hourly mean wind speed (m/s), r_h hourly rainfall total (mm), and ϕ_h hourly mean wind direction (degrees, north = 0°). Twelve orientations at 30° intervals were evaluated; the instrumented east-facing wall corresponds to $\Theta = 90°$. The measurement-period value was annualised as

$$I_{A,\mathrm{ann}}(\Theta) = I_{A,\mathrm{campaign}}(\Theta) \cdot \frac{365}{N_{\mathrm{days}}}, \tag{8}$$

where N_{days} is the duration of the weather subset in days. The values are reported as orientation-specific airfield indices. The roughness, topography, obstruction and wall factors used by BS EN ISO 15927-3:2009 to form a design wall index were not applied, because the analysis uses on-site weather observations to compare temporal wetting forcing during the monitoring period rather than to produce a façade design exposure value.

Dry-spell length τ_{dry} was defined as the number of consecutive days before the window centre with daily rainfall < 0.2 mm,

$$\tau_{\mathrm{dry}}(t) = \max\{k \in \mathbb{N} : R_{\mathrm{daily}}(t-i) < 0.2\,\mathrm{mm}, \ \forall\, i = 1, \ldots, k\}. \tag{9}$$

14-day converged U-values were stratified into *wet periods* ($\tau_{\mathrm{dry}} = 0$) and *dry spells* ($\tau_{\mathrm{dry}} \geq 2$ days) and compared with one-sided Mann–Whitney U tests. A one-sided test was used because the physical hypothesis is directional: liquid water is expected to raise, not lower, the apparent U-value [14,15,22]. The apparent U-value recovery rate was then estimated by ordinary-least-squares regression [47]

$$U = U_0 + k_{\mathrm{dry}}\, \tau_{\mathrm{dry}} + \epsilon, \tag{10}$$

in which $k_{\mathrm{dry}} < 0$ ($\mathrm{W\,m^{-2}\,K^{-1}\,day^{-1}}$) quantifies the wall's return towards its dry baseline. We emphasise that k_{dry} is an *effective* coefficient describing how the measured U-value declines with antecedent dry-spell length; it is not a direct measure of the rate of change of the actual liquid-water content of the wall, which was not measured. It is retained as a practical, measurable proxy for the drying-related recovery of the apparent thermal transmittance, and is interpreted as such throughout. Pearson correlations [48] between τ_{dry} and U, and between 14-day cumulative rainfall and U, were also computed.

2.5.5. Significance of features

To explain the residual scatter that remains after the duration and moisture analyses, twelve physically motivated and non-redundant features were extracted from each 14-day rolling window (Table 3). A Random-Forest regressor [49], as implemented in `scikit-learn` [50], was fitted separately at each location, using U as target on a 70/30 train/test split (200 trees, max depth 8, `min_samples_leaf = 5`, `random_state = 0`). These hyperparameters balance variance reduction (number of trees) against the overfitting risk on the modest sample size of ~100 14-day windows per location, and were confirmed by a coarse 5-fold cross-validation grid (depths 4–12, leaf sizes 3–10) that produced indistinguishable importance rankings. A Random-Forest model was preferred over linear regression because several candidate features are expected to interact non-linearly with U (for example, ΔT magnitude modulates the influence of heat-flux variability), and an ensemble of decision trees captures these interactions without imposing a parametric form.

Feature importance was assessed by *permutation importance* [49,51], which records the loss in predictive accuracy when a single feature is randomly shuffled. Permutation importance is preferred to the alternative impurity-based importance because the latter is biased towards high-cardinality features. Permutation importance of feature x_k was computed with $M = 30$ repeats on held-out samples as

$$I_k = \frac{1}{M} \sum_{m=1}^{M} \left[\mathrm{MSE}\big(\hat{U}(\tilde{X}^{(m,k)}), U\big) - \mathrm{MSE}\big(\hat{U}(X), U\big) \right], \tag{11}$$

where $\tilde{X}^{(m,k)}$ is the feature matrix with column k randomly permuted. The same feature set was used at both locations so that the rankings are directly comparable. Pearson and Spearman [52] correlations between each feature and U were also tabulated. To verify rank stability, the importance computation was repeated for ten random seeds; the outcome of this stability check is reported with the feature-importance results in Section 3.

3. Results

The four analyses are reported in the same order in which the methods were introduced, because each builds on the previous one. The length-of-monitoring analysis identifies the window length needed for a meaningful Average-Method estimate, and so justifies the 14-day base window used for everything that follows. The moisture analysis then explains why otherwise similar 14-day windows can give different answers. The starting-time analysis isolates the diurnal-phase effect within those same converged windows, and the feature-significance analysis ranks the experimental and environmental factors that account for the remaining scatter.

3.1. Raw measurements

Before the window-based analyses, Fig. 6 presents the raw 10-minute measurements over two representative four-week sub-periods, one in December 2025 and one in March 2026, chosen to show both early-winter and early-spring conditions. Three features of the data underpin the rest of the Results. First, the heat flux (panel a) is small and noisy: the baseline conductive flux is only of order 1 to 3 W/m^2, punctuated by sharp positive spikes that coincide with the daily heating cycle and, on a few occasions, by brief negative excursions during mild, solar-influenced afternoons. This low and intermittently sign-changing signal is the physical origin of the long averaging requirement quantified in Section 3. Second, the indoor air temperature (panel b, solid) cycles with the heating schedule while the outdoor temperature (panel b, dashed) follows weather synoptics, so the driving temperature difference ΔT (panel c) is highly non-stationary, mostly between 5 and 15 °C but dropping below the ISO 9869-1 $\Delta T = 5$ °C guidance floor during the mildest spells, for example the first week of December. Third, and most important for the credibility of the two-station design, the two locations track each

Table 3
Twelve physically motivated features used in the Random-Forest analysis of feature significance. Features are computed per 14-day rolling window.

Category	Feature	Definition
A. ΔT quality	dT_mean, dT_std, dT_range	Mean, std. dev. and range of $T_{in} - T_{out}$
B. Flux quality	q_abs_mean, q_std	Mean absolute and std. dev. of heat flux q
C. Thermal weather	Solar_mean, Wind_mean	Window-mean solar irradiance and wind speed
D. Moisture	Rain_total_mm, RH_mean	14-day cumulative rainfall and mean relative humidity
E. Indoor stability	Ti_std	Std. dev. of indoor air temperature
F. Surface temp.	Tse_mean	Mean external surface temperature
G. Diurnal cycling	q_diurnal_amp	Amplitude of diurnal q cycle (FFT first harmonic)

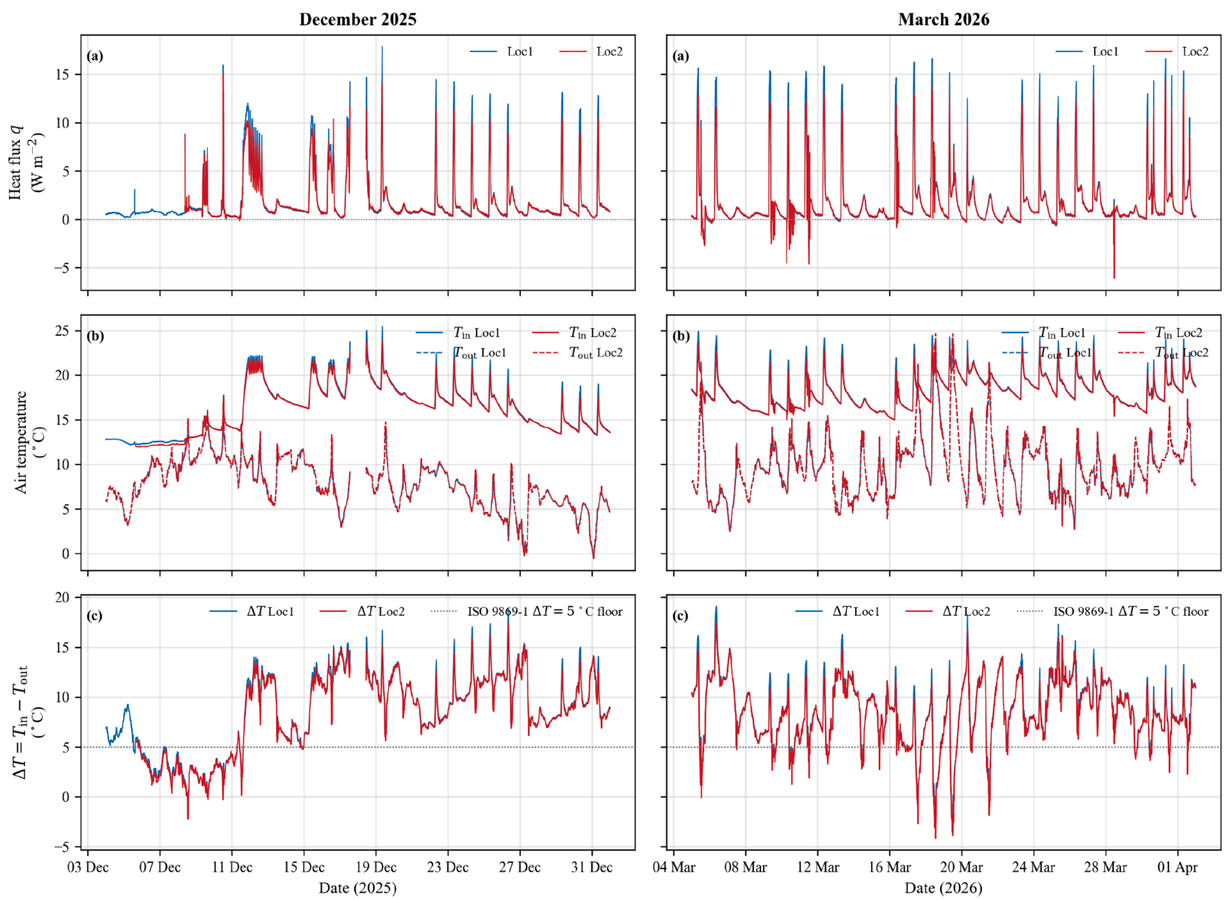

Fig. 6. Raw 10-minute measurements over two representative four-week sub-periods (left: December 2025; right: March 2026) at both sensor stations. (a) Heat-flux density q; (b) indoor (solid) and outdoor (dashed) air temperatures; (c) driving temperature difference $\Delta T = T_{in} - T_{out}$ with the ISO 9869-1 $\Delta T = 5$ °C guidance floor. The two stations track each other closely in both periods, indicating one-dimensional heat flow at both points..

other almost exactly in all three panels and in both periods, which confirms that the heat flow at both points is effectively one-dimensional and free of localised thermal bridging. The window-mean driving force is $\overline{\Delta T} = 8.5$ and 8.4 °C (Loc1, Loc2) in the December period and 8.7 and 8.5 °C in the March period, with indoor-temperature standard deviations of 2.7 and 2.6 °C in December and 1.8 and 1.7 °C in March; these values are used in the feature-significance analysis (Section 3) and in the ΔT and indoor-stability discussion.

3.2. *Length of monitoring*

The headline result is that the 72-hour minimum permitted by ISO 9869-1 does not give a stable U-value on this wall, while 14 days does, and 28 days is needed for what we term *certification-grade* convergence. Throughout this paper, “certification-grade” is defined specifically as a mean absolute error ≤5% *relative to the long-duration measured reference* U_{ref} (Eq. (6)); it is a precision/repeatability criterion describing how closely a short test reproduces the value obtained by long averaging with the same instrument on the same wall. It should not be confused with absolute measurement accuracy against the true U-value: Section 9 of ISO 9869-1 [10] estimates an expected total uncertainty of 14–28% for a generic in-situ HFM measurement, and the combined instrumentation uncertainty for the present wall is 4–5% per 14-day window (Section 2.4, Appendix A). The 5% convergence threshold used here is therefore a necessary, not a sufficient, condition for an accurate absolute U-value, and the design–measurement comparison is treated separately in Section 4. The same picture emerges whether one looks at the spread of measured U-values, the percentage error against a long-duration reference, or the standard’s own convergence rate.

Fig. 7 shows how the converged U-value distribution narrows as the window lengthens. At 3–7 days the boxes are wide (±0.06–0.09 W/(m^2 K)) and centred near, but often biased from, the

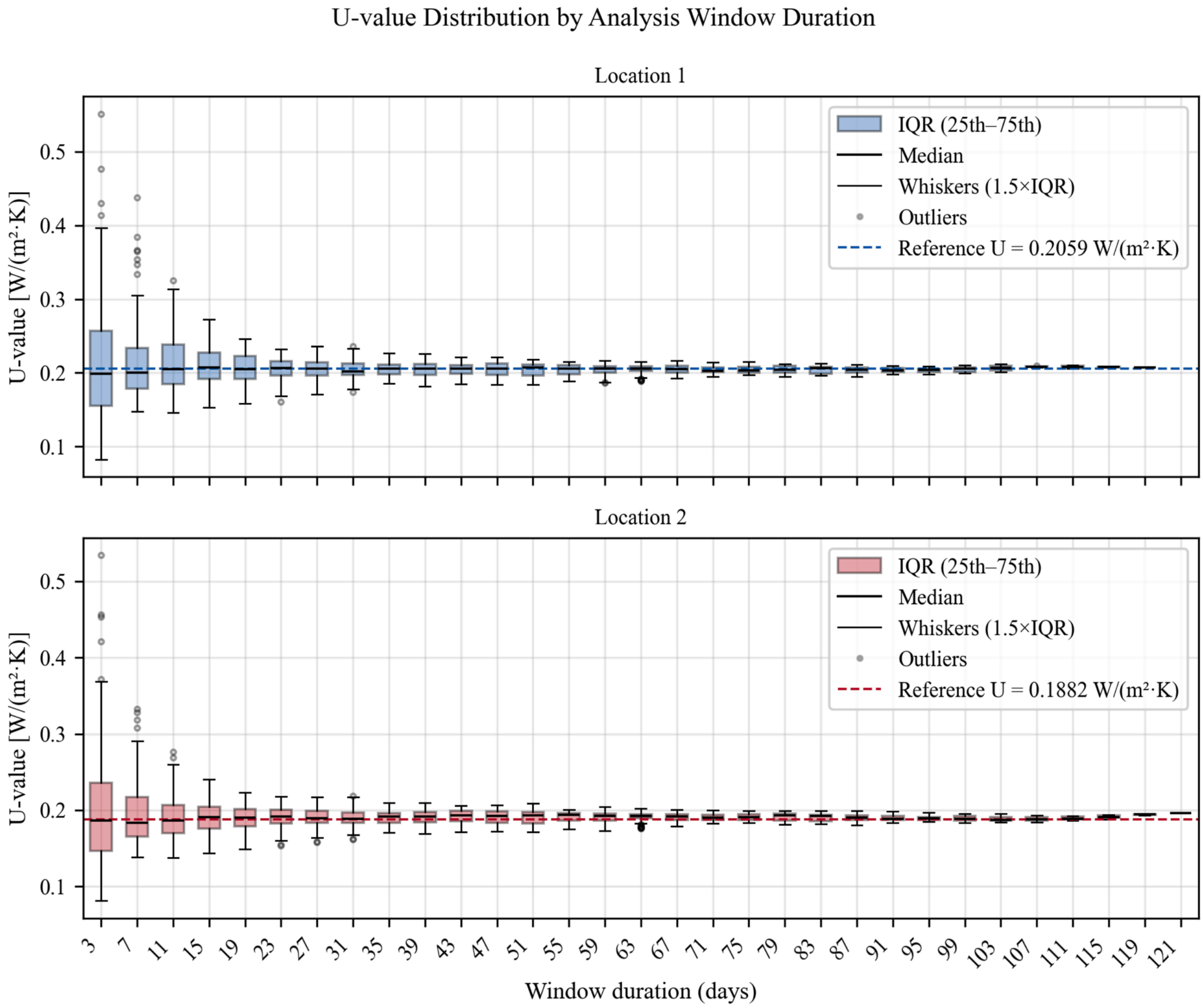

Fig. 7. Boxplots of converged U-values versus window duration. The dashed line marks the long-duration reference at each location (Eq. (6)). Scatter shrinks from ±0.06–0.09 W/(m^2 K) at 3–7 days to ±0.02 W/(m^2 K) at 14–21 days.

long-duration reference. By 14 days the interquartile range has contracted to about ±0.02 W/(m^2 K) and the medians sit on the reference line.

This narrowing translates directly into reduced error against the reference (Fig. 8, computed via Eq. (5)). The mean absolute error falls from about 32% at 72 h to 10–12% at 14 days, 5% at 28 days, and 3–4% at 42 days. To reach ≤3% mean absolute error at both locations, windows of approximately 65–70 days would be required, and ≤2% would require approximately 90 days, which is rarely practical in the field. Two features of Fig. 8 warrant explanation. First, the two location curves terminate at different window lengths (Loc1 just before 120 days, Loc2 just after): this is simply because the longest complete window that can be placed within the fixed 130-day record differs slightly between stations, since the two channels have marginally different start times and gap patterns after quality screening, so Loc2 retains one additional 24-hour step. Second, the Loc2 mean error ticks upward beyond ∼120 days; at these longest durations only one or two candidate windows remain (Section 2.5), so the statistic is dominated by a single window that happens to span the wettest sub-period, and the apparent rise reflects small-sample variability and residual moisture bias rather than a genuine loss of accuracy with longer averaging. This is also consistent with the ISO 9869-1 Clause 7.1 convergence requirements: such a wet-spanning window violates the 24-hour stability and first/last two-thirds criteria (Eqs. (3)–(4)), so it is not a convergence-compliant result in the first place. These longest-duration points are therefore plotted for completeness but should be read as indicative only.

The ISO 9869-1 convergence rate (Fig. 9) tells the same story from the standard's own perspective: only 5–10% of 72-hour windows simultaneously satisfy all three Clause 7.1 criteria, rising to 16–17% at 14 days, 31–41% at 30 days and 36–40% at 42 days. Throughout this range, the third criterion (Eq. (4)) is the binding constraint, in line with [13,16,17] on different wall typologies. The convergence rate rises to a broad plateau near 30–42 days and then declines for the longest windows (≳35–45 days in Fig. 9). This decline reflects the ISO 9869-1 Clause 7.1 convergence requirements as much as a sampling effect, and is not a physical reversal. First, a longer window accumulates more secular drift in R_D from moisture cycling and slowly varying boundary conditions, so a progressively larger fraction of long windows breach the 24-hour stability and first/last two-thirds tolerances (Eqs. (3)–(4)) and are rejected as non-converged, the third criterion remaining the binding one. Second, as the window grows towards the length of the record the number of independent placements falls sharply (from ∼183 at 28 days to a single window at 121 days), so each non-converged long window removes a larger fraction of the denominator, and any long window that straddles the wettest spell fails that third criterion through this same secular drift. The convergence rate at the longest durations is therefore estimated from very few windows and is correspondingly noisy; the practically relevant message is the monotonic rise up to ∼30 days.

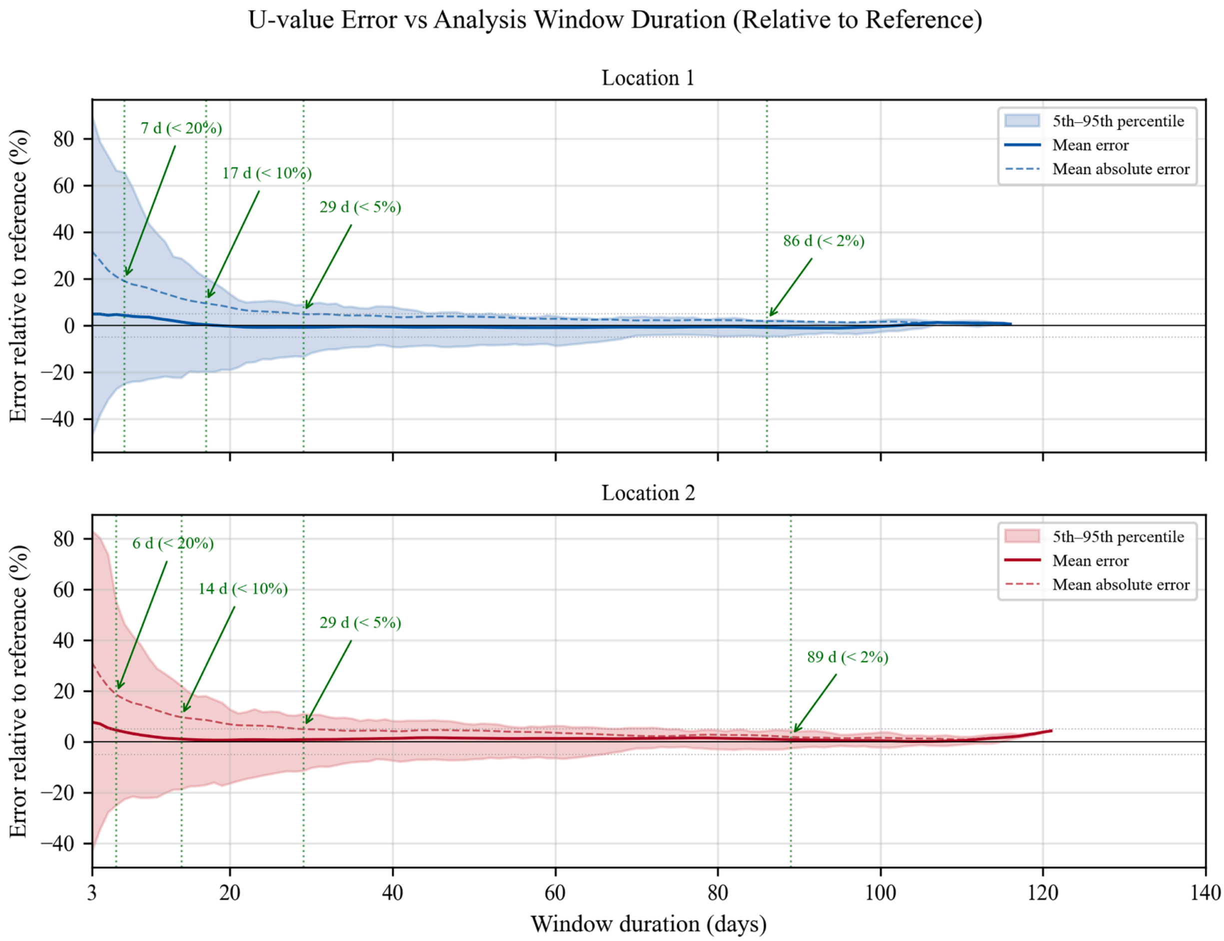

Fig. 8. Absolute percentage error $|\varepsilon|$ (Eq. (5)) versus window duration, with 5th–95th percentile bands. The ISO 9869-1 72-hour minimum leaves a central tendency of ~32% error on this post-retrofit low U-value wall.

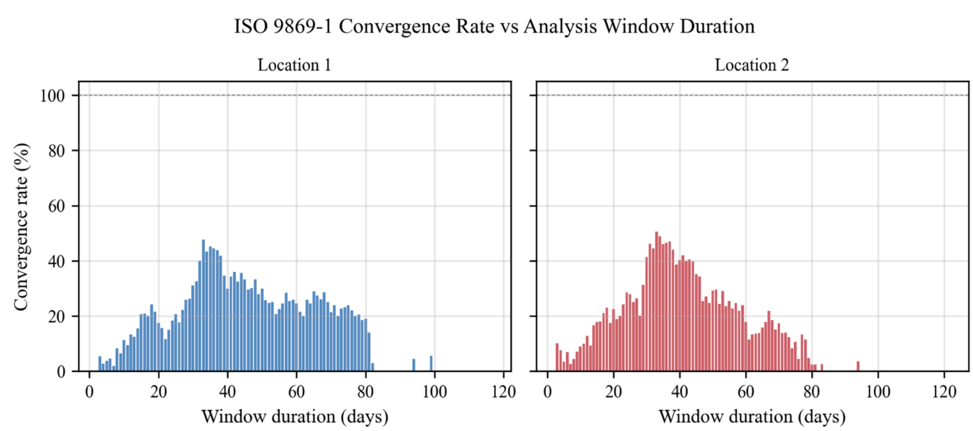

Fig. 9. ISO 9869-1 Clause 7.1 convergence rate (simultaneous satisfaction of Eqs. (2)–(4)) as a function of window duration for both locations. The 72-hour minimum permitted by the standard yields only 5–10% convergence; the rate reaches 31–41% only for windows of ≥30 days.

The rolling time series in Fig. 10 makes the same point in the time domain. Each panel plots, for one window length, the U-value of every rolling window against the date of the window centre: faint traces are all candidate windows at a 24-hour stride, coloured markers are the subset that satisfy the three Clause 7.1 convergence criteria, and the horizontal dashed lines are the long-duration reference U-values at the two locations (Loc1: 0.206, Loc2: 0.188 W/(m^2 K)). Short 3- and 7-day windows produce large excursions, ranging from below 0.10 to above 0.35 W/(m^2 K), that are not representative of the long-run wall performance. The 14- and 30-day curves are markedly more stable, and converge onto the reference dashed lines as the window lengthens. The two locations also track each other closely, confirming internal

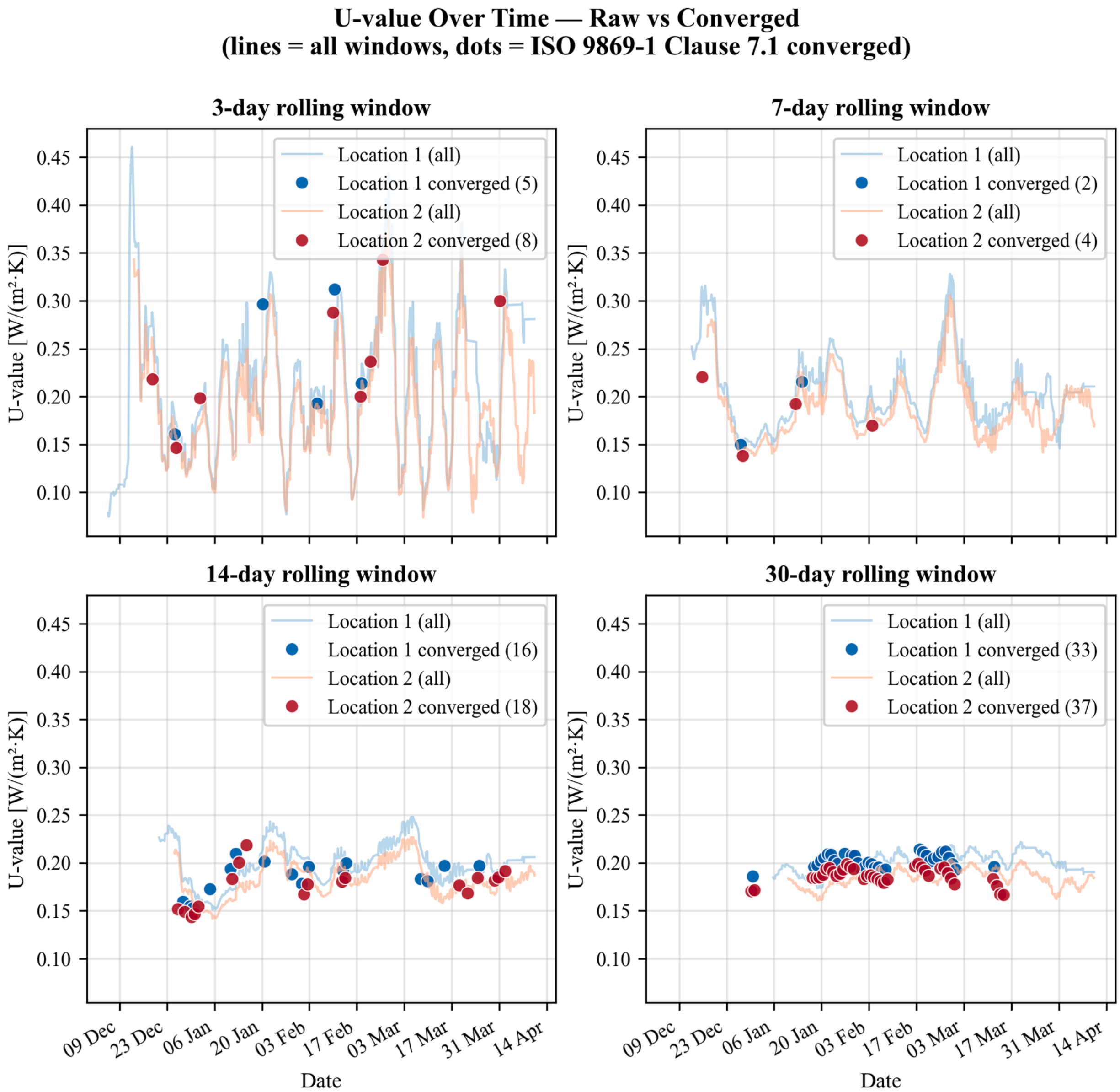

Fig. 10. Rolling U-value time series at four window durations. Faint traces show raw rolling values at 24-hour step; coloured markers mark ISO 9869-1 Clause 7.1 converged windows (Eqs. (2)–(4)). Reference U-values (Loc1: 0.206, Loc2: 0.188 W/(m^2 K)). Convergence rates: 3-d 5–10%, 7-d 2–3%, 14-d 16–17%, 30-d 31–41%.

consistency between the two stations. A mid-period rise visible in the 30-day panel coincides with the wettest sub-period and motivates the moisture analysis that follows.

The physical reason that 72 hours is insufficient on this wall is essentially one of signal-to-noise. With a design U-value of 0.18 W/(m^2 K) and a typical winter ΔT of about 10 °C, the heat flux through the wall is only ~1.8 W/m^2, an order of magnitude smaller than for an uninsulated solid-brick wall. Sensor noise, diurnal cycling and moisture transients therefore make up a much larger fraction of the measured signal, and longer averaging is required to recover a stable mean. A 14-day floor brings the mean error below 12%, and a 28-day window is needed for ≤5%; this duration ladder is consistent with the ranges previously reported for low U-value façades by [8,18], for heavyweight masonry by [14,15], and for post-retrofit walls by Farmer et al. [20], Wakili et al. [26]. With the 14-day base window now justified, the moisture analysis can ask why otherwise similar 14-day windows give different answers.

3.3. Moisture impact and wind-driven rain

The length-of-monitoring analysis showed that 14-day windows are stable enough to compare like with like. Three sub-questions can therefore be asked in turn: how exposed is the east-facing wall to wind-driven rain, do high-WDR episodes coincide with high measured U-values, and how much does antecedent dry weather change the measured value? These three questions, the climatic exposure of the east facade, the coupling between wind-driven rain and measured U-value, and the wet-versus-dry stratification, are examined in turn below.

The site context is summarised in Fig. 11, in which the on-site wind rose is overlaid on satellite imagery. Prevailing winds are WSW/SW and NE; the test wall is partially sheltered to the west by neighbouring buildings, but exposed to easterly fetch. Occasional easterly events therefore align with the normal to the instrumented wall and dominate direct WDR loading on the measurement surface.

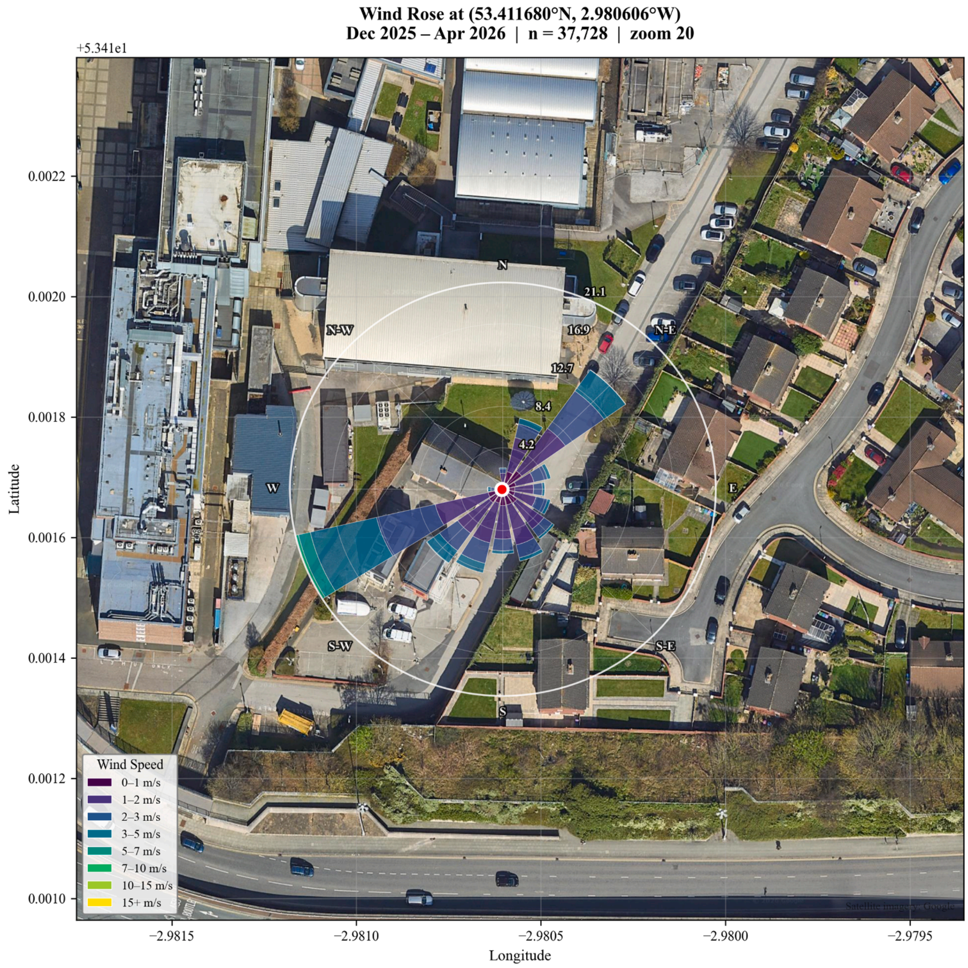

Fig. 11. Wind rose over satellite imagery at the measurement site, with OpenStreetMap building footprints providing sheltering context. Prevailing winds are WSW/SW and NE.

Applying Eq. (7) to the on-site weather data gives the orientation-specific airfield index in Fig. 12. The east-facing measurement wall has a monitoring-period index of 35.8 mm, equivalent to $I_{A,\text{ann}} = 101$ mm/yr when annualised by Eq. (8). The most exposed orientation overall is SSW (210°) at 162 mm/yr, with WSW only slightly lower; the east face is therefore not the worst-case orientation, but it does intercept direct WDR during easterly storms that, as the next sub-section shows, coincide with peaks in measured U-value.

The orientation-weighted index proves more informative than total rainfall: only 35.8 mm of the 250.0 mm monitoring-period rainfall (14.3%) actually drove onto the east face. Fig. 13 overlays this east-facing WDR series on the converged U-values, and the two track each other closely. The heaviest WDR episodes in late January–early February coincide with the highest measured U-values in both 14- and 30-day windows, providing direct evidence of moisture–U-value coupling. The apparent reduction in measured U-value in early February, even though parts of this period are shaded as "wet", has two related explanations. The first is the dry week around 3 February: a run of dry days allowed the east face to lose moisture, lowering the apparent U-value. The second is a question of how the rolling values are plotted. Each marker is the U-value of a multi-week window plotted against the date of the window *centre*, so it is not the U-value of that single day; it reflects the antecedent conditions over the whole window, and both the wetting and the subsequent drying of the wall act with a lag of several days to a week, consistent with the recovery rate reported below. A short "wet" label on the omnidirectional rainfall record therefore does not translate into an immediate rise in the windowed U-value. A third and smaller contribution is orientation: the wet shading in Fig. 13 is based on total (omnidirectional) rolling rainfall, whereas the bias-inducing quantity is the east-facing WDR index (Eq. (7)), and part of the early-February rain fell under westerly and south-westerly winds onto the sheltered face rather than the instrumented east wall. These effects are why the orientation-specific WDR index, and not total rainfall, is used as the moisture driver throughout this study.

The coupling becomes quantitative when the 14-day windows are stratified by antecedent dry-spell length (Eq. (9)) and compared in Fig. 14. The mean U-value falls by **15.1%** at Loc1 ($0.1934 \rightarrow 0.1642$ W/(m^2 K), $p = 0.001$) and **13.1%** at Loc2 ($0.1836 \rightarrow 0.1596$ W/(m^2 K), $p = 0.015$) as one moves from wet ($\tau_{\text{dry}} = 0$) to dry-spell ($\tau_{\text{dry}} \geq 2$ days) windows. Fitting Eq. (10) gives drying-rate estimates of $k_{\text{dry}} = -6.5$ mW/(m^2 K)/day at Loc1 and -3.9 mW/(m^2 K)/day at Loc2, and the underlying scatter is strongly negative (Pearson $r = -0.86$, $p < 0.001$ at Loc1; $r = -0.64$, $p = 0.005$ at Loc2). Roughly one week of dry weather is therefore needed before the wall returns towards its dry baseline.

The wet–dry difference is larger than the instrumentation uncertainty estimated in Section 2.4, so it cannot be attributed to sensor accuracy. It is important to note that this 13–16% figure is a *relative, within-instrument* difference between two sets of windows measured by the *same* sensor on the *same* wall (wet minus dry). The 14–28% total uncertainty quoted in Section 9 of ISO 9869-1 is an *absolute* accuracy figure that is dominated by systematic terms (plate calibration, contact resistance, fixed thermal-bridging offset) which are common to the wet and dry groups and therefore cancel almost entirely in the wet-minus-

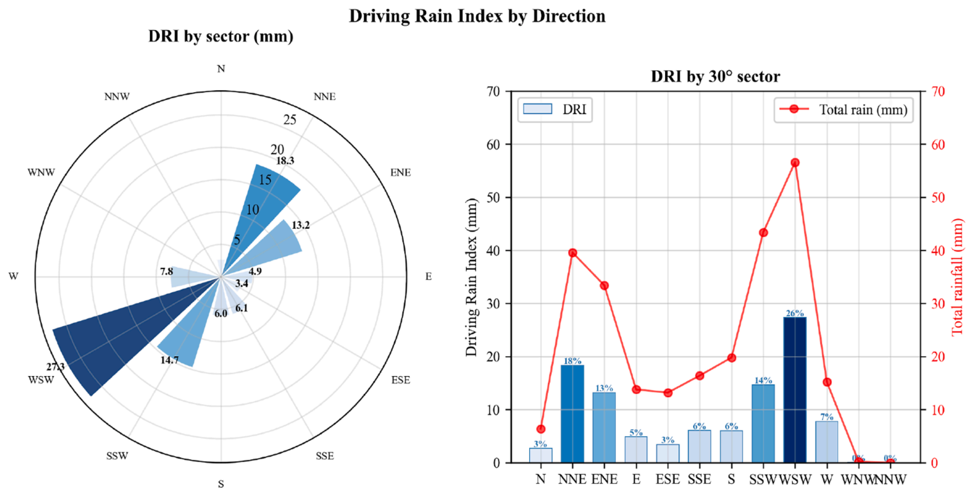

Fig. 12. BS EN ISO 15927-3:2009 orientation-specific airfield index by 30° wall orientation (Eq. (7)). The east-facing measurement wall has a monitoring-period index of 35.8 mm and an annualised index of 101 mm/yr; the maximum annualised orientation is SSW (210°) at 162 mm/yr.

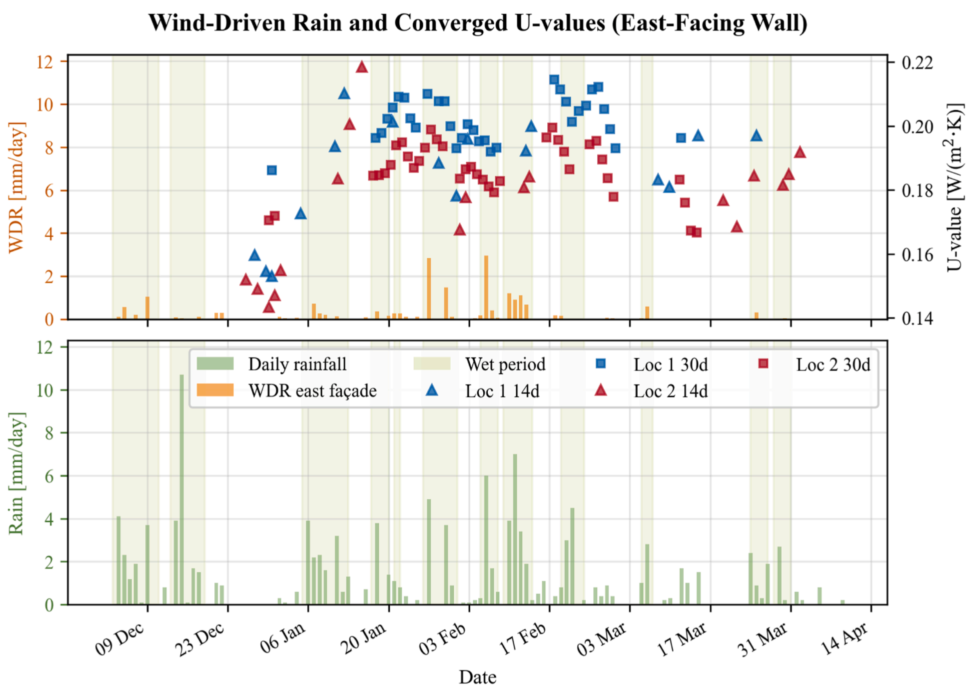

Fig. 13. WDR on the east wall (top panel, orange bars, Eq. (7)) overlaid with converged U-value markers (14-day triangles, 30-day squares), and daily total rainfall (bottom panel). Shading marks wet periods (rolling 3-day rainfall >60th percentile). High-WDR episodes align with peaks in measured U-value.

dry comparison. What remains is the much smaller random/repeatability component (~4–5% per 14-day window, Section 2.4), against which a 13–16% wet–dry shift is statistically significant ($p \leq 0.015$). The wet–dry effect is thus a genuine signal and not an artefact of the standard's headline absolute-uncertainty band. It also turns a long-recognised physical effect [14,15,22] into specific guidance: on a post-retrofit wall with EWI in a maritime-temperate climate, the reported U-value is sensitive to antecedent weather at the order of ±15%, and an explicit dry-spell pre-conditioning period is therefore warranted. We are careful here not to overstate the physical claim: the analysis quantifies the *effect of antecedent dry-spell length on the measured U-value*, not the absolute moisture content of the wall, which was not measured. Expressions describing moisture as "quantified" should be read in this measurement-response sense throughout. The presence of an EWI layer adds further hygrothermal complexity, since moisture absorbed by the brick-slip cladding or trapped at the EWI–brick interface may take longer to redistribute than in unclad masonry, which reinforces the case for pre-conditioning.

3.4. *Starting time*

With duration and moisture as the two large effects accounted for, the starting-time analysis tests whether a smaller, scheduling-related effect remains. The relevant statistics are presented below. Mean converged U-value is essentially flat as a function of start hour for every duration tested (Fig. 15): the largest spread across the 24-hour clock is about 2% of the mean, and it falls below 1% once windows are 7 days or longer. Formal tests in Table 4 confirm this: night-started and

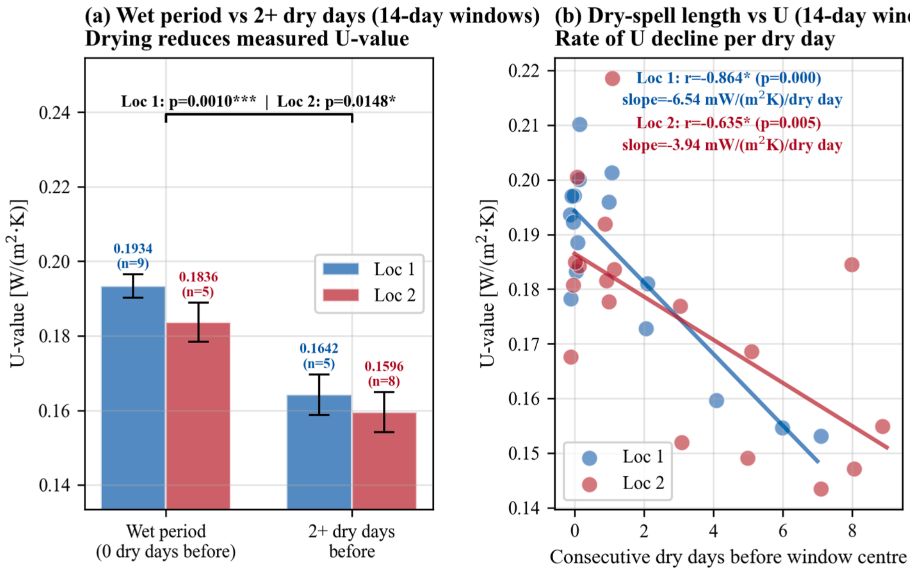

Fig. 14. (a) Mean 14-day converged U-values for wet periods ($\tau_{dry} = 0$) vs $\tau_{dry} \geq 2$-day dry spells, with standard errors; Mann–Whitney significance annotated. (b) Dry-spell length (Eq. (9)) vs U-value, with linear regressions (Eq. (10)) and Pearson correlations. Drying rates are $k_{dry} = -6.5$ and -3.9 mW/(m^2 K)/day for Loc1 and Loc2.

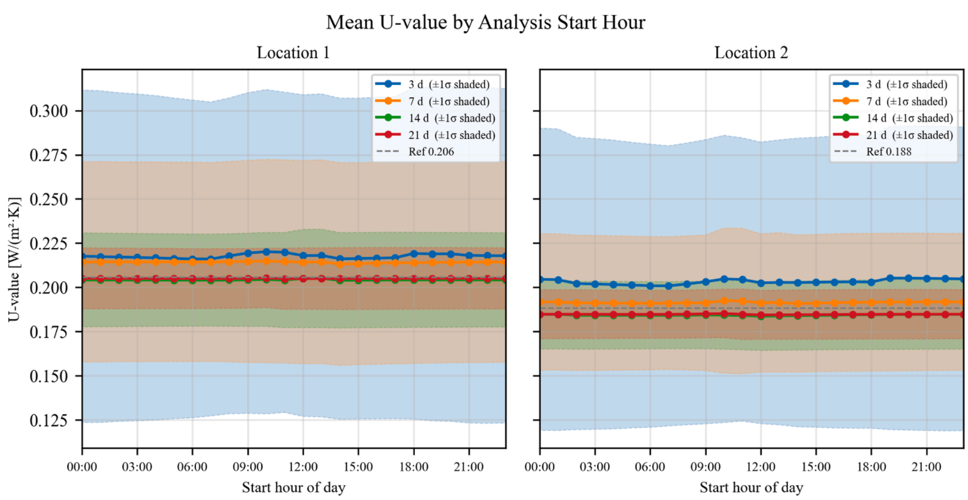

Fig. 15. Mean converged U-value as a function of test start hour for four durations at both locations. The 24-hour profile is flat within the confidence bands, showing that for the Average Method with $D \geq 3$ days the start time is not a material control.

Table 4
Starting time statistics (night 18:00–06:00 vs day 06:00–18:00). All differences are statistically non-significant.

Duration	ΔU (%)	Mann–Whitney p	Cohen's d	Kruskal-Wallis p
3 days	<0.30	>0.8	<0.01	>0.9
5 days	<0.20	>0.7	<0.01	>0.9
7 days	<0.10	>0.7	<0.01	>0.9
10 days	<0.20	>0.9	<0.01	>0.9
14 days	<0.20	>0.7	<0.02	>0.9
21 days	<0.05	>0.9	<0.01	>0.9

day-started windows are statistically indistinguishable at every duration (Mann–Whitney $p > 0.7$, Cohen's $d < 0.02$), and the Kruskal–Wallis omnibus test across all 24 hours is also non-significant.

This null result is practically useful. For Average-Method tests of three days or more on a low U-value retrofit wall, scheduling tests for a particular hour of the day brings no measurable benefit, because the integration in Eq. (1) averages out diurnal forcing regardless of phase.

3.5. Significance of other features

The final question is what controls the residual scatter that remains after duration is fixed at 14 days and the moisture effect has been characterised. Random-Forest permutation importance (Eq. (11)) gives a clear answer: measurement-protocol variables dominate, and the same five factors dominate at both locations even though the exact rank order differs slightly. Permutation importance quantifies how much the model's prediction accuracy (here, mean squared error on the U-value) degrades when the values of a given feature are randomly shuffled, thereby breaking its relationship with the target; a higher value means the model relies more heavily on that feature and the feature explains more of the window-to-window scatter in measured U-value. Values sum to approximately 1 across all twelve features. The enumeration below is ordered by the Loc1 ranking (which also matches the rank obtained when the two locations are pooled); the Loc2 permutation-importance values are reported alongside for the same features, and their order at Loc2 differs (at Loc2 the leading features are `dT_std`, `q_abs_mean` and `Ti_std`, with `dT_mean` ranked fourth at 0.080). This is the intended message of

Fig. 16: the *same* five protocol features dominate at both stations, but their internal order is not identical, so the apparently low Loc2 value of `dT_mean` is consistent with, not contradictory to, Fig. 16. In descending order of importance these are

1. `dT_mean`, mean indoor–outdoor temperature difference $\overline{\Delta T} = \overline{T_{in} - T_{out}}$ (°C) over the 14-day window: driving-force magnitude; larger $\overline{\Delta T}$ raises the signal-to-noise ratio of Eq. (1). Permutation importance: Loc1 0.284, Loc2 0.080.
2. `dT_std`, standard deviation of the 10-minute ΔT series (°C) over the window: driving-force variability, a proxy for thermal non-stationarity of the boundary conditions. Permutation importance: Loc1 0.160, Loc2 0.276.
3. `q_abs_mean`, mean absolute heat flux $\overline{|q|}$ ($W\,m^{-2}$) over the window: direct signal-to-noise indicator for the heat-flux measurement; low values indicate near-zero net heat transfer and high relative sensor noise. Permutation importance: Loc1 0.114, Loc2 0.213.
4. `Ti_std`, standard deviation of the indoor air temperature T_{in} (°C) over the window: indoor temperature stability; large excursions drive additional thermal-mass charging/discharging that biases the cumulative average. Permutation importance: Loc1 0.107, Loc2 0.171.
5. `q_std`, standard deviation of the 10-minute heat-flux series ($W\,m^{-2}$) over the window: heat-flux variability, reflecting combined sensor noise and transient conduction not captured by the mean. Permutation importance: Loc1 0.055, Loc2 0.069.

Fig. 16 ranks all twelve features at each location. This ranking is robust: repeating the permutation-importance computation for ten random seeds left the top-five order unchanged at both locations, with importance values varying by less than ±15%. The weather variables sit lower in raw importance, but as the moisture analysis already showed they act *indirectly* through moisture accumulation: 14-day cumulative rainfall correlates positively with U-value (Pearson $r = +0.41$ at Loc1, $p < 10^{-4}$; $r = +0.20$ at Loc2, $p = 0.04$). There is no contradiction between weather appearing low in the Random-Forest importance and moisture having a significant effect in the wet–dry analysis. The two analyses ask different questions on different timescales. The wet–dry stratification compares groups of windows separated by antecedent dry-spell length and isolates the moisture signal directly; the Random-Forest, by contrast, ranks how much each feature explains the *window-to-window scatter that remains after the 14-day duration is already fixed*, and on that residual the instantaneous protocol variables (driving-force magnitude and stability, heat-flux signal-to-noise, indoor stability) vary more sharply from window to window than the slowly varying, partly collinear weather aggregates. Moisture's influence is real but is expressed through the cumulative-rainfall feature (with the positive correlations quoted above) and through its slow modulation of the heat-flux signal, rather than through a single dominant instantaneous weather feature. In other words, weather sets the slow drift while protocol quality sets the fast scatter.

The bivariate scatter plots in Fig. 17 show that the signs of these effects are consistent with physical intuition, provided they are read as *measurement-artefact* signatures rather than as changes in the true fabric U-value (which is a property of the wall and does not itself depend on ΔT). A larger $\overline{\Delta T}$ is associated with a *lower* measured U-value because it raises the heat-flux signal relative to the fixed sensor-noise floor, reducing the positive noise-rectification bias in the ratio of sums in Eq. (1); greater ΔT variability and indoor instability are associated with a *higher* measured U-value because they degrade the quasi-steady assumption and leave un-cancelled thermal-storage terms in the numerator; and a higher mean heat-flux magnitude is weakly associated with a higher measured U-value once the ΔT regime is fixed. We stress that these associations describe how the *estimator* responds to measurement conditions, not how the wall itself changes. The linear fits in Fig. 17 are deliberately shown only to indicate the *sign* and rough consistency of each relationship between the two stations; the scatter is wide and the linear coefficients of determination are modest, which is exactly why a non-linear Random-Forest, rather than linear regression, was used for the importance ranking. The linear lines should therefore be read as indicative trends, not as well-determined predictive models. Taken together, the four analyses give a coherent operational picture for this wall: choose a window of at least 14 days, avoid the week immediately after heavy WDR, do not worry about the exact start hour, and ensure ΔT and indoor temperature are as large and steady as possible.

4. Discussion

4.1. Revising the 72-hour minimum for high-performance retrofit fabric

The convergence and error curves (Figs. 9 and 8) suggest that the blanket 72-hour minimum in ISO 9869-1 is not appropriate for high-performance post-retrofit envelopes. For the wall studied, 14 days appears to be a reasonable lower bound for an indicative U-value (errors ≤12%), 28 days is required for certification-grade (≤5%) results, and approximately 65–70 days is required for ≤3% mean absolute error; even 42-day windows still carry ~4% mean absolute error.

The underlying reason is essentially physical. A design U-value of 0.18 $W/(m^2\,K)$ produces a steady-state heat flux of only ~1.8 W/m^2 under a 10 °C temperature difference, compared with ~15 W/m^2 for an uninsulated solid-brick wall at U-value ≈ 1.5 $W/(m^2\,K)$. At such low flux levels, the sensor noise floor (~0.5 W/m^2 for typical HFM plates), diurnal cycling and solar-driven transients on the east façade together represent a substantial fraction of the measured signal, so longer averaging periods are needed to recover a stable mean. This signal-to-noise argument is consistent with Gaspar et al. [18], who found that façades with U-value < 0.5 $W/(m^2\,K)$ required $\Delta T > 19$ °C or extended durations to achieve 5% accuracy, and with O'Hegarty et al. [8], who documented extreme scatter for U-value < 0.2 $W/(m^2\,K)$.

The third Clause 7.1 criterion (Eq. (4)) [10] also appears to be the binding constraint throughout. As already noted by Lee et al. [16], Choi and Ko [17], tests that are "accurate" in the sense that $U \approx U_{ref}$ may nevertheless fail this criterion because of slow secular drift in the environmental boundary conditions. A revised criterion that recognises the thermal time constant of the wall, for example a duration set to at least three thermal time constants rather than a fixed 72 hours, would be more physically grounded and would particularly benefit the growing stock of insulated retrofit walls whose time constants are much longer than the 72-hour floor was originally designed for. Two caveats apply to a time-constant-based rule. First, the thermal time constant $\tau = RC$ requires the layer-by-layer thermal resistances and heat capacities to be known, which is often not the case for as-built retrofit walls where the exact cavity-fill density, EWI thickness and brick-slip build-up may be uncertain or undocumented. Where the layer make-up is unknown, τ can instead be estimated empirically from the measured data, for example from the decay time of the heat-flux autocorrelation or from the e-folding time of the step response after a sharp ΔT change, so that the rule remains usable without a full material specification. Second, the present numbers translate to other walls through the same physics rather than as fixed day counts: the required duration scales with the wall's thermal time constant and inversely with the achievable signal-to-noise ratio (itself set by the U-value and the available ΔT). A heavier or better-insulated wall, or one tested at smaller ΔT, will therefore need proportionally longer than the 14/28-day figures found here, whereas a lighter or higher-U-value wall tested at large ΔT may converge sooner. The practical recommendation is thus to express the floor as "≥3 thermal time constants, and not less than 14 days for fabric below ~0.3 $W/(m^2\,K)$", so that the rule adapts to other wall types rather than fixing a single universal duration. For more poorly insulated, higher-U-value walls, where the heat-flux signal is far larger and convergence is correspondingly rapid, the existing 72-hour ISO 9869-1 minimum generally remains valid, and the extended durations recommended here apply specifically to low U-value fabric; further long-term monitoring would

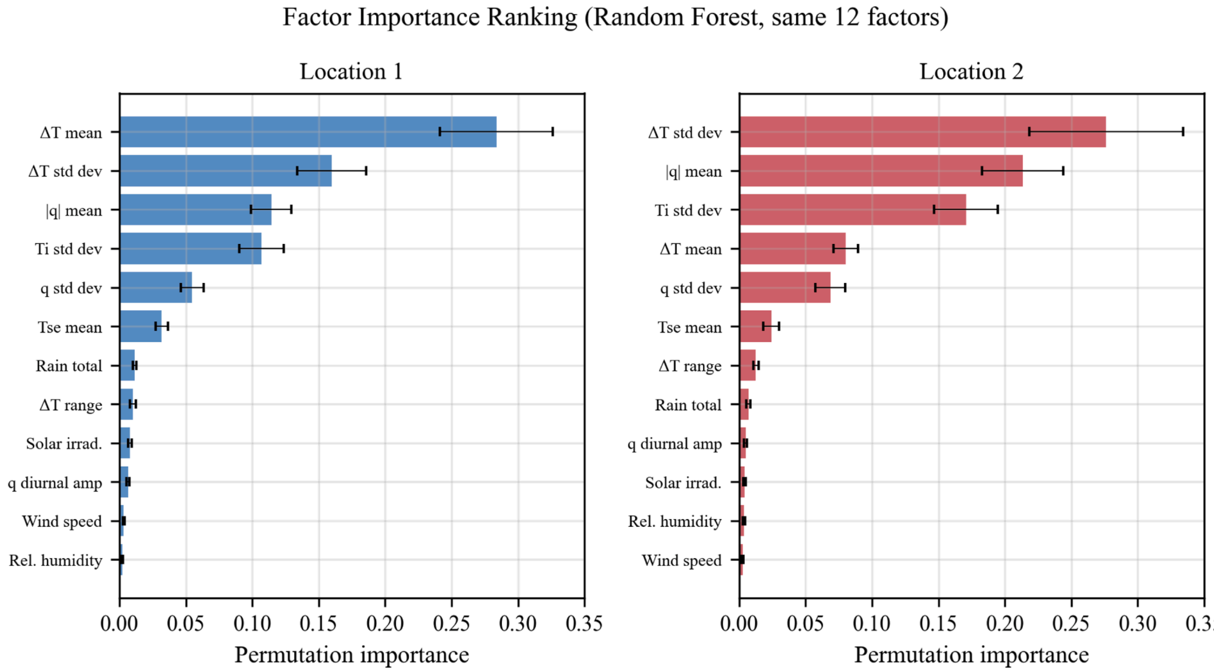

Fig. 16. Random-Forest permutation importance (Eq. (11)) of the 12 factors controlling 14-day U-value at each location (200 trees, max depth 8, 30 repeats). Each panel is ranked independently by its own permutation importance, so the factor order differs between the two locations. The same five factors dominate across locations, confirming that measurement-protocol sensitivities dominate over site-specific effects.

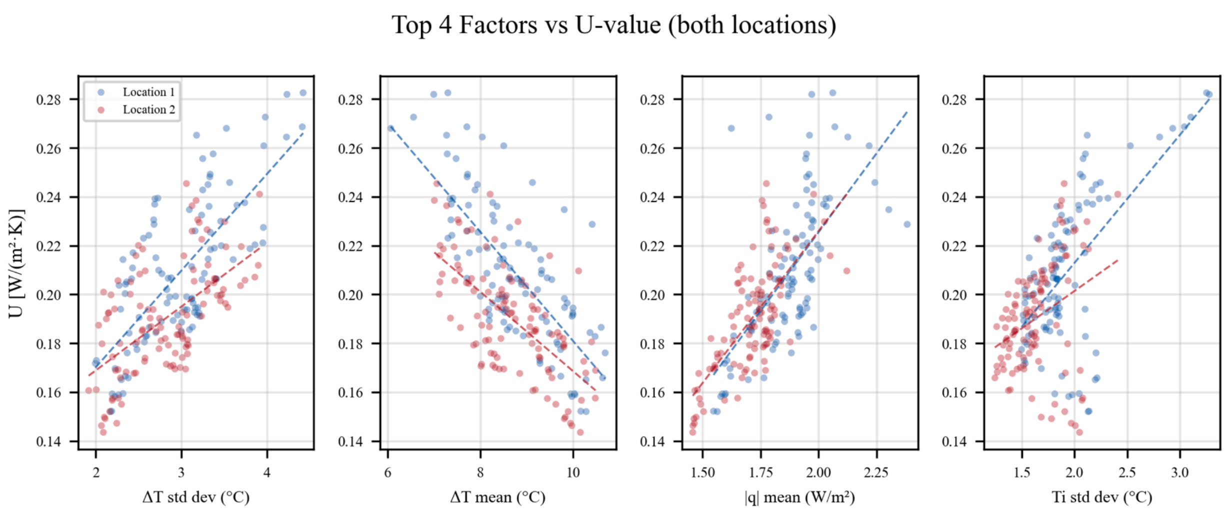

Fig. 17. Scatter of U-value against the four top-ranked factors, with both locations overlaid. Linear regressions confirm consistent signs and similar slopes between locations.

be needed to test how weather and moisture affect the measured U-value of such higher-U-value walls in the way examined here.

This duration ladder (14 / 28 / 65–70 days) is broadly consistent with the ranges reported by Gaspar et al. [19] in their review of HFM minimum-test-duration criteria, by Lucchi [14], Genova and Fatta [15] for historic heavyweight masonry, and by Farmer et al. [20], Wakili et al. [26] for post-retrofit constructions. The present study extends the evidence base from those earlier works to a modern UK cavity-brick retrofit with EWI, a construction type that has received relatively less attention than either historic masonry or continental nZEB envelopes.

4.2. Relation to dynamic and active measurement methods

The analysis above stays within the framework of the steady-state Average Method of ISO 9869-1, because that is the method in routine use for retrofit sign-off and the one whose 72-hour floor is under examination here. It is nonetheless worth setting the results against the family of dynamic and active alternatives, several of which are explicitly designed to shorten the test or to reduce the ΔT sensitivity that dominates low U-value measurements. Dynamic (system-identification) post-processing of the same heat-flux and temperature records, for example the ISO 9869-1 dynamic method, ARX/RC grey-box models, the Anderlind regression, and Bayesian or Kalman-filter state estimators [28,30,31,53], explicitly models the wall's thermal storage instead of waiting for it to average out, and can in favourable cases return a converged U-value from one to two weeks of data, sometimes less. These methods directly target the slow-drift mechanism that we found makes the third Clause 7.1 criterion binding, and are therefore a natural complement to the duration ladder proposed here: where a 28-day steady-state test is impractical, a dynamic re-analysis of a 10–14-day record may recover comparable accuracy, at the cost of greater modelling complexity and sensitivity to model-order and prior choices. Active methods that impose a controlled boundary condition (guarded or unguarded hot-box style heating, the Excitation-Response or heat-pulse techniques, and infrared-thermography-assisted approaches [30,39]) raise the heat-flux signal artificially and so improve the signal-to-noise ratio that is the root cause of the long durations reported here; they are powerful for low U-value fabric but require additional equipment, controlled access and, often, unoccupied conditions, which limits their use as a routine in-situ quality-assurance hold point during occupied retrofit delivery. Our findings should therefore be read as establishing a defensible *passive, steady-state* baseline, against which the additional cost and complexity of dynamic post-processing or active excitation can be justified when test duration must be compressed or when ΔT is unavoidably small. A direct, like-for-like comparison of the dynamic and Average methods on this wall is the subject of ongoing work: as shown in the outdoor and indoor installation photographs (Figs. 3 and 4), a second pair of heat-flux plates was installed on both flanking walls for this purpose, and a dedicated companion paper will report the dynamic-versus-Average comparison using the extended record, including summer data, where the small driving temperature difference is most challenging; that comparison is kept separate here to keep the present paper to a manageable length.

4.3. Starting time irrelevance

The null result of the starting time analysis (Fig. 15, Table 4) carries two practical implications. First, practitioners working with the Average Method over ≥3-day windows need not schedule tests for specific hours of the day, and measurements may be initiated whenever access permits. Second, this is not, however, a licence to ignore solar exposure: the mechanism that renders starting time irrelevant is the time-integration inherent in Eq. (1), which averages across many diurnal cycles, and the recommendation of Evangelisti et al. [29] to avoid south-facing or directly sun-exposed walls therefore remains valid. The *mean* solar gain over the test shifts the apparent U-value, while the *phase* of the start does not.

4.4. Moisture as the dominant external control

The 13–16% wet-vs-dry gap (Fig. 14) and the ~4–6 mW/(m^2 K)/day drying rate together provide a quantitative link between WDR and the measured U-value. The underlying mechanism is reasonably clear: liquid water raises the apparent thermal conductivity of the insulation and masonry layers, and latent-heat transfer adds to the measured heat flux [14,15,21,22]. When this is set against an east-facing BS EN ISO 15927-3:2009 annualised airfield index of 101 mm/yr, the effect cannot be regarded as a marginal correction, because a field monitoring that begins in the week following an easterly storm could be biased by 13–16% even before any instrumentation error is taken into account. For an EWI-clad wall, moisture pathways are also more complex than for exposed masonry, in that WDR may be absorbed by the brick-slip render, trapped at the insulation–masonry interface, or channelled through fixings [26]. The drying rates measured here (4–6 mW/(m^2 K)/day) should therefore be interpreted as effective wall-system rates that integrate across all of these layers.

It would seem reasonable that ISO 9869-1 reporting should include the WDR and dry-spell context at the time of measurement. A short climatic-exposure annex, following the BS EN ISO 15927-3:2009 framework used here, could complement the existing requirement [10] to report measurement start/end times and logging intervals.

4.5. Interpreting the design–measurement difference

The reference U-values of 0.206 W/(m^2 K) (Loc1) and 0.188 W/(m^2 K) (Loc2) exceed the design value of 0.18 W/(m^2 K) [38] by 14.4% and 4.4%, respectively. In absolute terms, however, the differences are only +0.026 and +0.008 W/(m^2 K), with a mean absolute difference of +0.017 W/(m^2 K). For a low U-value fabric this is, on the present evidence, a fairly modest deviation, and one that should not be interpreted in the same way as a large absolute heat-loss penalty. A number of factors may contribute to this difference, including local thermal bridging at EWI fixings, workmanship quality at the cavity-fill boundary, moisture uptake during the first heating season after retrofit, and the inherent uncertainty in steady-state calculations that assume homogeneous material properties [31]. The ~10% difference observed between the two sensor locations, which are mounted on the same wall under nominally identical boundary conditions, also suggests that spatial variability on a single façade can be of comparable magnitude to the relative design–measurement difference itself.

From a standards and retrofit quality-assurance standpoint, a purely relative performance-gap threshold can therefore over-penalise high-performance fabric, since a small absolute difference is divided by a small design U-value. ISO 9869-1 reporting would arguably be more informative if design–measurement agreement were assessed against both a relative tolerance and an absolute U-value tolerance. For low U-value walls, the result might reasonably be considered within tolerance if either criterion is satisfied, for example if

$$\left|U_{\mathrm{meas}} - U_{\mathrm{design}}\right| \le \Delta U_{\mathrm{abs}} \quad \text{or} \quad \frac{\left|U_{\mathrm{meas}} - U_{\mathrm{design}}\right|}{U_{\mathrm{design}}} \le \delta_{\mathrm{rel}}. \tag{12}$$

Such a dual criterion would preserve relative performance-gap reporting for higher U-value elements while allowing absolute U-value differences to provide the more meaningful pass/fail basis for low U-value fabric. On the present evidence, indicative threshold values of $\Delta U_{\mathrm{abs}} \approx 0.03$ W/(m^2 K) and $\delta_{\mathrm{rel}} \approx 15\%$ would have classified both sensor locations as within tolerance against the BS EN ISO 6946 design value, while still flagging the genuinely under-performing solid-wall-insulation installations reported in the ECO4/GBIS audits [5] where measured U-values exceed design by 0.5 W/(m^2 K) or more. These suggested values are illustrative; calibration against a multi-site dataset would be needed before adoption in a standard.

The ~10% difference between the two sensor locations is, however, based on $n = 2$ pseudo-replicates on the same wall and should be read as indicative of within-wall variability rather than as a precise estimate of facade-scale heterogeneity.

4.6. What matters most: protocol, not weather

The feature-significance analysis ranks measurement-protocol variables (ΔT quality, heat-flux signal strength, indoor stability) above weather variables in raw importance (Fig. 16). This ordering is consistent with the moisture-impact analysis, since weather acts indirectly through moisture accumulation over 14-day horizons while protocol variables determine the signal-to-noise ratio in Eq. (1). A reliable U-value therefore appears to require four conditions: ΔT should be maximised, ideally above 10 °C in line with Lee et al. [16], Ficco et al. [36], Bienvenido-Huertas et al. [54]; indoor temperature should be kept as stable as practicable; the week immediately following heavy WDR should be avoided; and the test should integrate over at least 14 days, and preferably 28.

4.7. Implications for warm homes delivery and retrofit quality assurance

These findings position in-situ U-value measurement as a practical quality-assurance hold point for large-scale retrofit, and not solely as a retrospective research tool. The Warm Homes Plan links delivery scale with consumer protection, standards reform, oversight, and the use of in-use thermal-performance evidence to tailor measures, identify hidden defects and improve installation quality [3]. The ECO4/GBIS response indicates why this matters in practice: visual inspection, certification and paperwork did not in themselves prevent high rates of major non-compliance in some solid-wall-insulation installations, and consequently required audit, suspension and remediation at no cost to affected households [4,5].

For practitioners, the actionable message of the present analysis is that a 72-hour ISO 9869-1 test cannot be regarded as a defensible sign-off test for this class of low U-value wall. Commissioning specifications would more reasonably require a 14-day minimum for indicative diagnosis, 28 days where results are used for certification or dispute resolution, explicit reporting of the preceding dry spell and orientation-specific WDR index, and the reporting of both absolute and relative design–measurement differences. For housing providers and local authorities delivering area-based programmes, these requirements could be applied as a risk-based sample across archetypes, elevations, contractors and installation batches, with failed or moisture-biased tests triggering re-test, intrusive inspection or targeted remediation before final payment is released. For policy makers and scheme administrators, the implication is that monitoring should be embedded as a scheme rule rather than left to optional post-occupancy research, since measured fabric performance can complement PAS 2035, PAS 2030 and TrustMark compliance checks, protect households from ineffective insulation, and provide early evidence on whether Warm Homes Plan funding is in fact producing real heat-loss reductions.

This study does not imply that every dwelling in the UK requires 130-day HFM monitoring. Rather, it indicates the minimum measurement physics that should inform proportionate scheme design: low U-value walls need longer integration, an explicit moisture context, and a tolerance framework that distinguishes material underperformance from small absolute differences. Embedding these checks would help avoid two opposite failures in retrofit quality assurance, namely accepting poorly performing insulation because the measurement was too short or too wet, and overstating a modest low U-value discrepancy because it has been judged only as a percentage.

4.8. Limitations

Four limitations merit explicit mention. First, the present results are based on a single wall with two heat-flux sensors in a single climate and over a single (winter–early spring) season, and the two sensors are pseudo-replicates on the same façade rather than independent samples. The numerical thresholds reported here (14 / 28 / 65–70 days, the 13–16% wet-vs-dry effect, and the indicative $\Delta U_{\mathrm{abs}} \approx 0.03$ W/(m^2 K) tolerance) should therefore be read as a case-study calibration of the underlying signal-to-noise and thermal-time-constant arguments, rather than as universal numbers; multi-site replication across wall typologies, orientations and climates is required for generalisation. Lee et al. [25] and Nardi and Lucchi [39] note that convergence behaviour is sensitive to climate zone, wall thermal mass and internal gain patterns. Second, the reference U-value was taken as the median of the longest converged windows rather than as an independently measured laboratory U-value; an independent guarded-hot-box reference for the same wall build-up would tighten the absolute-error analysis, although the first-half/second-half U_{ref} check (Section 2.4) and the ±0.02 W/(m^2 K) spread at ≥14 days lend support to the choice. Third, the 130-day monitoring spans winter and early spring only; extending the monitoring through summer would allow the seasonal dependence of convergence to be characterised more fully [24], which is of particular interest for low U-value walls where the ΔT driving force diminishes in warmer months. Because the present campaign did not include summer data, it cannot itself evidence warm-season behaviour, and the duration figures above should not be extrapolated to summer without further measurement. The governing constraint is the driving temperature difference rather than the calendar season: ISO 9869-1 already requires ΔT to remain above about 5 °C, and on a low U-value wall the small ~2 W/m^2 signal makes this floor the binding limitation. In hot-summer climates, such as the Italian setting examined by Alongi et al. [55], ΔT frequently falls towards a few degrees and can reverse during the day; their parametric heat-flow study found that wall thermal-property estimates degrade markedly under small and reversing temperature differences, so summer Average-Method testing is correspondingly unreliable there. In mild or cool maritime climates, by contrast, a sufficient indoor-to-outdoor temperature difference can often be maintained or induced through summer, in which case Average-Method testing remains feasible provided ΔT is kept above the ISO floor. Where a reliable ΔT cannot be sustained, a dynamic or active method (above) is preferable to a longer steady-state test. Fourth, the wet-vs-dry stratification uses total daily rainfall as the dryness indicator; a stricter test stratifying by orientation-specific WDR (Eq. (7)) would more directly isolate east-façade wetting and is a natural extension once additional sites are available.

5. Conclusions

A 130-day, two-sensor monitoring study on a post-retrofit east-facing wall (cavity brick + insulation + EWI + brick slip; design U-value = 0.18 W/(m^2 K)) in Liverpool, UK, analysed at 10-minute resolution over 13,687 candidate windows across both locations, provides the evidence base for adapting ISO 9869-1 to high-performance retrofit fabric. The findings cohere around a single physical theme: a low U-value wall produces a heat-flux signal so small (~2 W/m^2) that test *duration*, antecedent *moisture* and the relative-versus-absolute framing of the *design gap* all become first-order issues, whereas the start hour does not. The four conclusions below should be read as consequences of that one signal-to-noise argument rather than as independent results. Four conclusions follow:

1. **The 72-hour standard minimum is inadequate** for low U-value retrofit walls; mean absolute error at 72 h is ~32%. A 14-day duration floor should be adopted for indicative results, 28 days for ±5% certification-grade accuracy, and approximately 65–70 days for ≤3% mean absolute error (42-day windows still carry ~4% mean absolute error). The physical basis is the low signal-to-noise ratio inherent in high-performance envelopes: at U-value = 0.2 W/(m^2 K) and ΔT = 10 °C, the heat flux is only ~2 W/m^2, making long averaging essential.
2. **The starting time of an Average-Method test is statistically irrelevant** once the window is ≥3 days ($p > 0.7$, $d < 0.02$). Scheduling tests for specific hours of the day confers no benefit.
3. **Moisture is the dominant external control**: wet periods inflate U-value by 13–16% relative to dry spells, with 4–6 mW/(m^2 K) recovered per consecutive dry day. The east-facing BS EN ISO 15927-3:2009 airfield index is 35.8 mm for the monitoring period, or 101 mm/yr when annualised, and measurement tests following heavy WDR require explicit pre-conditioning.
4. **The design–measurement difference is marginal in absolute terms**: measured U-value exceeds the 0.18 W/(m^2 K) design value by +0.008 to 0.026 W/(m^2 K), even though the relative difference reaches 4–14%. For low U-value fabric, absolute U-value differences should be reported alongside percentages and may provide the more appropriate compliance tolerance. Protocol quality (ΔT magnitude, heat-flux signal strength, indoor temperature stability) sets window-to-window scatter; weather acts indirectly through 14-day moisture accumulation (Pearson r = +0.20 to +0.41, $p < 0.05$).

The results support augmenting ISO 9869-1 Annex B with (i) a wall-type-dependent duration floor calibrated to the target U-value and local ΔT regime, (ii) a mandatory dry-spell pre-conditioning period scaled to the local orientation-specific WDR index, and (iii) a dual tolerance for judging design–measurement agreement, under which either an absolute U-value limit or a relative percentage limit can be satisfied. Adapting the standard in these three respects would bring reported Average-Method U-values closer to true as-built performance, avoid overstating small absolute differences in low U-value fabric, and give practitioners, policy makers and housing providers a defensible monitoring route for the Warm Homes Plan and successor retrofit schemes [3,6,8,9].

CRediT authorship contribution statement

Hu Du: Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Alexander Pilkington:** Data curation.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used Claude Opus 4.7 and GPT-5.5 in order to assist with code generation. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

The equipment cost was jointly supported by Liverpool John Moores University and Wienerberger UK Ltd. The building retrofit was sponsored by Wienerberger UK Ltd. Neither funder had any role in the study design, data collection, analysis, interpretation, writing of the manuscript or the decision to submit it for publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to acknowledge Wienerberger UK Ltd. for sponsoring the building retrofit; Liverpool John Moores University and Wienerberger UK Ltd. for sponsoring the equipment

Data availability

The data that support the findings, including 10-minute sensor data (`merged_10min.csv`), 5-minute weather data (`weather_5min_clean.csv`) and Python analysis scripts, will be deposited in the LJMU Data Repository at https://opendata.ljmu.ac.uk/ following a embargo from the date of publication to allow further publications. During the embargo period, the data are available from the corresponding author on reasonable request.

Appendix A. Combined standard uncertainty of the Average-Method U-value

The Average-Method U-value (Eq. (1)) is a ratio of the time-summed heat flux to the time-summed temperature difference,

$$U = \frac{\sum_{j=1}^{N} q_j}{\sum_{j=1}^{N} \left(T_{\mathrm{in},j} - T_{\mathrm{out},j}\right)} = \frac{\overline{q}}{\overline{\Delta T}}. \tag{A.1}$$

Following the GUM [41] framework for propagation of uncertainty through a function $U = f(\overline{q}, \overline{\Delta T})$, and treating the mean heat flux and the mean temperature difference as the two dominant input quantities, the combined standard uncertainty $u_c(U)$ is obtained from the first-order (relative) propagation law as

$$\left(\frac{u_c(U)}{U}\right)^2 = \left(\frac{u(\overline{q})}{\overline{q}}\right)^2 + \left(\frac{u(\overline{\Delta T})}{\overline{\Delta T}}\right)^2 - 2\,\frac{u(\overline{q}, \overline{\Delta T})}{\overline{q}\,\overline{\Delta T}}, \tag{A.2}$$

where $u(\overline{q})$ and $u(\overline{\Delta T})$ are the standard uncertainties of the window-mean heat flux and temperature difference and $u(\overline{q}, \overline{\Delta T})$ is their covariance. For the present analysis the heat-flux and temperature channels are calibrated independently, so the covariance term is taken as negligible and Eq. (A.2) reduces to a root-sum-square of the two relative terms. Each input uncertainty combines a systematic (calibration) component and a random (averaging) component,

$$u(\overline{q}) = \sqrt{u_{\mathrm{cal},q}^2 + \frac{s_q^2}{N_{\mathrm{eff}}}}, \qquad u(\overline{\Delta T}) = \sqrt{u_{\mathrm{cal},\Delta T}^2 + \frac{s_{\Delta T}^2}{N_{\mathrm{eff}}}}, \tag{A.3}$$

in which s_q and $s_{\Delta T}$ are the standard deviations of the 10-minute samples within the window, and N_{eff} is the effective number of independent samples after accounting for autocorrelation (estimated from the integral autocorrelation time, $N_{\mathrm{eff}} = N\,\Delta t/(2\tau_{\mathrm{int}})$). Using the manufacturer calibration figures of Section 2.4 (±3% for the heat-flux plate and ±0.1–0.2 K per temperature channel), a 14-day window with $\overline{\Delta T} \approx 8$ °C yields a relative systematic contribution of ≈3% from the plate and ≈2% from $\overline{\Delta T}$, with the random terms contributing a further 1–2% after autocorrelation correction, giving a combined standard uncertainty of approximately 4–5% per 14-day window. The expanded uncertainty at the 95% level ($k = 2$) is therefore of order 8–10%, which remains within the 14–28% total band quoted in Section 9 of ISO 9869-1 [10] and against which the 13–16% wet–dry effect (Section 3.3) is distinguishable because the dominant systematic terms cancel in the wet-minus-dry comparison. In plain terms, a single 14-day Average-Method U-value from this wall is limited by the instruments to about ±5% (roughly ±0.01 W/(m^2 K) on a 0.2 W/(m^2 K) wall), widening to about ±8–10% at the stricter 95% confidence level; a difference in the reported U-value smaller than this band cannot be separated from sensor noise, whereas the larger moisture and design–measurement effects discussed in the main text can be.

References

[1] M. Teni, H. Krstić, P. Kosiński, Review and comparison of current experimental approaches for in-situ measurements of building walls thermal transmittance, Energy Build. 203 (2019) 109417.
[2] L. Evangelisti, E. De Cristo, S. Monteleone, C. Guattari, P. Gori, T. de Rubeis, et al., Development and validation of a non-contact method for U-value estimation in heritage and existing buildings, Result. Eng. 28 (2025) 107276.
[3] Department for Energy Security & Net Zero, Warm Homes Plan, Policy Paper, DESNZ, London, 2026. Available from: https://www.gov.uk/government/publications/warm-homes-plan/warm-homes-plan-html.
[4] Department for Energy Security & Net Zero, Action Taken to Protect Households with Poor-Quality Insulation, DESNZ, London, 2025, (Press release). https://www.gov.uk/government/news/action-taken-to-protect-households-with-poor-quality-insulation.
[5] Department for Energy Security & Net Zero, Ofgem, Solid Wall Insulation Installed Under ECO4 and GBIS: Statistical Audit Results. Research and Analysis DESNZ, London, 2025. https://www.gov.uk/government/publications/solid-wall-insulation-installed-under-eco4-and-gbis-statistical-audit-results/.
[6] A. Marshall, R. Fitton, W. Swan, D. Farmer, D. Johnston, M. Benjaber et al., Domestic building fabric performance: closing the gap between the in situ measured and modelled performance, Energy Build. 150 (2017) 307–317.
[7] F. Li, A. Smith, P. Biddulph, I. Hamilton, R. Lowe, A. Mavrogianni, et al., Solid-wall U-values: heat flux measurements compared with standard assumptions, Build. Res. Inf. 43 (2) (2015) 238–252.
[8] R. O'Hegarty, O. Kinnane, D. Lennon, S. Colclough, In-situ U-value monitoring of highly insulated building envelopes: review and experimental investigation, Energy Build. 252 (2021) 111447.
[9] H. Er-Retby, M. Charai, M. Mghazli, M. El Mankibi, M. Benzaazoua, In-situ and data-driven methods for evaluating building energy performance in the operational phase: a critical review, Energy Strategy Rev. 63 (2026) 102038.
[10] ISO. BS ISO 9869-1:2014 Thermal insulation — building elements — in-situ measurement of thermal resistance and thermal transmittance. Part 1: heat flow meter method. International Organization for Standardization, 2014.
[11] K. Gaspar, M. Casals, M. Gangolells, A comparison of standardized calculation methods for in situ measurements of façades U-value, Energy Build. 130 (2016) 592–599.
[12] I. Atsonios, I. Mandilaras, D. Kontogeorgos, M. Founti, A comparative assessment of the standardized methods for the in-situ measurement of the thermal resistance of building walls, Energy Build. 154 (2017) 198–206.
[13] L. Evangelisti, C. Guattari, E. De Lieto Vollaro, F. Asdrubali, Convergence criteria analysis for thermal conductance measurements of building walls: a case study, Case Stud. Therm. Eng. 49 (2023) 103249.

[14] E. Lucchi, Thermal transmittance of historical brick masonries: a comparison among standard data, analytical calculation procedures, and in situ heat flow meter measurements, Energy Build. 134 (2017) 171–184.
[15] E. Genova, G. Fatta, The thermal performances of historic masonry: in-situ measurements of thermal conductance on calcarenite stone walls in Palermo, Energy Build. 168 (2018) 180–189.
[16] Y.-J. Lee, J.-H. Moon, D.-S. Choi, M.-J. Ko, Application of the heat flow meter method and extended average method to improve the accuracy of in situ U-value estimations of highly insulated building walls, Sustainability 16 (13) (2024) 5687.
[17] D. Choi, M. Ko, Analysis of convergence characteristics of average method regulated by ISO 9869-1 for evaluating in situ thermal resistance and thermal transmittance of opaque exterior walls, Energies 12 (10) (2019) 1989.
[18] K. Gaspar, M. Casals, M. Gangolells, In situ measurement of façades with a low U-value: avoiding deviations, Energy Build. 170 (2018) 61–73.
[19] K. Gaspar, M. Casals, M. Gangolells, Review of criteria for determining HFM minimum test duration, Energy Build. 176 (2018) 360–370.
[20] D. Farmer, C. Gorse, W. Swan, R. Fitton, M. Brooke-Peat, D. Miles-Shenton, et al., Measuring thermal performance in steady-state conditions at each stage of a full fabric retrofit to a solid wall dwelling, Energy Build. 156 (2017) 404–414.
[21] L. Caruso, V. Buhagiar, G. Caruana Smith, M. Larcher, S. Buhagiar, D. Bottino Leone, Hygrothermal performance of an innovative resource efficient composite masonry unit with embedded biobased insulation, Energy Build. 344 (2025) 116015.
[22] S. Flores Larsen, M. Hongn, N. Castro, S. González, Comparison of four in-situ methods for the determination of walls thermal resistance in free-running buildings with alternating heat flux in different seasons, Constr. Build. Mater. 224 (2019) 455–473.
[23] C. Li, H. Hou, D. Yang, A field framework for U-values and dynamic response in prefabricated rural dwellings: theory–measurement gaps and ranking reversal, Energy Build. 359 (2026) 117339.
[24] S. Juricic, J. Goffart, S. Rouchier, A. Foucquier, N. Cellier, G. Fraisse, Influence of natural weather variability on the thermal characterisation of a building envelope, Appl. Energy 288 (2021) 116582.
[25] Y.-J. Lee, J.-H. Moon, D.-S. Choi, M.-J. Ko, Influences of average temperature difference and measurement period on estimation of in situ thermal transmittance of building exterior walls using the average method of ISO 9869-1, Energies 17 (5) (2024) 1177.
[26] K.G. Wakili, B. Binder, M. Zimmermann, C. Tanner, Efficiency verification of a combination of high performance and conventional insulation layers in retrofitting a 130-year old building, Energy Build. 82 (2014) 237–242.
[27] A. Rodler, S. Guernouti, M. Musy, Bayesian inference method for in situ thermal conductivity and heat capacity identification: comparison to ISO standard, Constr. Build. Mater. 224 (2019) 126–139.
[28] V. Gori, C. Elwell, Estimation of thermophysical properties from in-situ measurements in all seasons: quantifying and reducing errors using dynamic grey-box methods, Energy Build. 167 (2018) 290–300.
[29] L. Evangelisti, C. Guattari, R. De Lieto Vollaro, F. Asdrubali, A methodological approach for heat-flow meter data post-processing under different climatic conditions and wall orientations, Energy Build. 223 (2020) 110216.
[30] A. Rasooli, L. Itard, In-situ characterization of walls' thermal resistance: an extension to the ISO 9869 standard method, Energy Build. 179 (2018) 374–383.
[31] P. Biddulph, V. Gori, C.A. Elwell, C. Scott, C. Rye, R. Lowe, et al., Inferring the thermal resistance and effective thermal mass of a wall using frequent temperature and heat flux measurements, Energy Build. 78 (2014) 10–16.
[32] A. Ahmad, M. Maslehuddin, L.M. Al-Hadhrami, In situ measurement of thermal transmittance and thermal resistance of hollow reinforced precast concrete walls, Energy Build. 84 (2014) 132–141.
[33] S.-H. Kim, J.-H. Kim, H.-G. Jeong, K.-D. Song, Reliability field test of the air–surface temperature ratio method for in situ measurement of U-values, Energies 11 (4) (2018) 803.
[34] D. Bienvenido-Huertas, C. Rubio-Bellido, J.L. Pérez-Ordóñez, M.J. Oliveira, Automation and optimization of in-situ assessment of wall thermal transmittance using a random forest algorithm, Build Env. 168 (2020) 106479.
[35] L. Evangelisti, C. Guattari, F. Asdrubali, Influence of heating systems on thermal transmittance evaluations: simulations, experimental measurements and data post-processing, Energy Build. 168 (2018) 180–190.
[36] G. Ficco, F. Iannetta, E. Ianniello, F. d'Ambrosio Alfano, M. Dell'Isola, U-value in situ measurement for energy diagnosis of existing buildings, Energy Build. 104 (2015) 108–121.
[37] S. Bergero, A. Cavalletti, A. Chiari, C. Marafioti, Experimental on-site measurement of thermal conductance by means of heat flow meter applied to nanocomposite thermal insulating mortar coating, J. Phys.: Conf. Ser. 2893 (2024) 012041.
[38] ISO. BS EN ISO 6946:2017 Building components and building elements — thermal resistance and thermal transmittance — calculation methods. International Organization for Standardization, 2017.
[39] I. Nardi, E. Lucchi, In situ thermal transmittance assessment of the building envelope: practical advice and outlooks for standard and innovative procedures, Energies 16 (8) (2023) 3319.
[40] B. Mobaraki, F.J. Castilla Pascual, A. Martínez García, M.Á. Mellado Mascaraque, B. Frutos Vázquez, C. Alonso, Studying the impacts of test condition and nonoptimal positioning of the sensors on the accuracy of the in-situ U-value measurement, Heliyon 9 (7) (2023) e17282.
[41] JCGM, Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement (GUM), JCGM 100:2008, Joint Committee for Guides in Metrology, 2008.
[42] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. 18 (1) (1947) 50–60.
[43] W. Kruskal, W. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. 47 (260) (1952) 583–621.
[44] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd ed., 1988.
[45] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat Soc. Ser. B 57 (1) (1995) 289–300.
[46] ISO. BS EN ISO 15927-3:2009, Hygrothermal performance of buildings — Calculation and presentation of climatic data — part 3: calculation of a driving rain index for vertical surfaces from hourly wind and rain data. International Organization for Standardization, 2009.
[47] D. C. Montgomery, E. A. Peck, G. G. Vining, Introduction to Linear Regression Analysis, John Wiley & Sons, Hoboken, NJ, 5th ed., 2012.
[48] K. Pearson, Note on regression and inheritance in the case of two parents, Proc. R. Soc. Lond. 58 (1895) 240–242.
[49] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[51] A. Altmann, L. Tološi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347.
[52] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1) (1904) 72–101.
[53] A. H. Deconinck, S. Roels, Comparison of characterisation methods determining the thermal resistance of building components from onsite measurements, Energy Build. 130 (2016) 309–320.
[54] D. Bienvenido-Huertas, J. Moyano, D. Marín, R. Fresco-Contreras, Review of in situ methods for assessing the thermal transmittance of walls, Renew. Sustain. Energy Rev. 102 (2019) 356–371.
[55] A. Alongi, L. Sala, A. Angelotti, L. Mazzarella, In situ measurement of wall thermal properties: parametric investigation of the heat flow methods through virtual experiments data, Energies 16 (10) (2023) 4247.