Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Establishing expected performance for residential air-to-water heat pumps under real-world conditions

Ugne Potthoff

ugne.potthoff@mtec.ethz.ch

ETH Zurich https://orcid.org/0009-0007-8495-2971

Constanze Bongs

Karlsruhe University of Applied Sciences,

Thorsten Staake

University of Bamberg https://orcid.org/0000-0003-1399-4676

Felix Wortmann

University of St. Gallen https://orcid.org/0000-0001-5034-2023

Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11529036/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. T.S. declares that he is a supervisory board member of Hoval, which manufactures and sells heating systems, among other products and services. U.P., C.B. and F.W. declare no competing interests. The authors' views do not necessarily represent those of the funding or collaborating institutions. Funding partners reviewed the manuscript only for confidential or proprietary information and had no role in study design, data analysis, interpretation, or the decision to publish.

Establishing expected performance for residential air-to-water heat pumps under real-world conditions

Abstract

As heat pumps become increasingly prevalent in residential buildings, discrepancies between expected and achieved efficiency can increase operating costs and undermine confidence in this technology. Yet, determining whether individual installations perform as expected remains challenging because efficiency strongly depends on operating conditions. Conventional assessments based on absolute efficiency thresholds cannot distinguish demanding operating conditions from additional performance shortfalls, while scalable methods for establishing installation-specific expectations are lacking. Here, we develop an approach to estimate expected heat pump efficiency under actual operating conditions, combining thermodynamic relationships, product-specific laboratory data, and field measurements. Validated using 176 heat pumps, the approach achieves a median absolute percentage error of 7.03%. Applying the method to 10,087 installations, we identify 5.1% with both low absolute efficiency and substantial deviations from expected performance. Systems falling below 80% of expected performance exhibit 2.3 times the incidence of subsequent recorded service contacts, whereas conventional absolute efficiency thresholds show no substantial difference between groups. These findings establish a scalable basis for assessing whether individual heat pumps perform as expected, enabling targeted performance monitoring and supporting greater transparency and efficiency in residential heating.

1 Introduction

Buildings account for approximately 30% of global final energy consumption, with space and water heating responsible for around half of building energy use [1]. Decarbonizing this demand requires both rapid heat pump deployment and efficient operation of installed systems. Heat pumps already supply around 12% of global heating demand in buildings, while annual global sales represent approximately 108 GW of heating capacity [2, 3]. Operating efficiency directly affects electricity consumption, costs, and the economic benefits of heat pump investments. In a study of 297 preselected Swiss households, on-site optimization reduced electricity consumption by an average of 15.2% in approximately half of the inspected installations [4]. However, such expert-led assessments cannot readily be scaled to millions of installations, particularly given shortages of qualified personnel [3]. Reliable methods for evaluating installed heat pumps and prioritizing improvements are therefore increasingly important to identify underperforming systems, enable targeted operational improvements, and guide future system design.

The measured efficiency of a heat pump depends not only on its laboratory-tested performance but also on the conditions of its installation, particularly outdoor and supply temperatures, domestic-hot-water demand, and system operation [5–7]. A seasonal performance factor (SPF) of 2.8, for example, may be expected in a high-temperature radiator system but indicate underperformance for the same product supplying low-temperature floor heating. A low SPF therefore does not necessarily indicate a problem with the heat pump itself.

Existing methods address different aspects of heat pump efficiency. Standardized product ratings, such as seasonal coefficient of performance (SCOP) under EN 14825, characterize performance under prescribed conditions, while VDI 4650 estimates annual efficiency using application-specific design assumptions [8, 9]. Field-monitoring studies document achieved performance [10, 11], and short-term monitoring methods estimate seasonal efficiency from limited observations [12]. A previous field-based approach estimated space-heating SCOP under standardized climatic conditions by fitting heating-curve and COP models to individual installations [13]. Although this enabled comparisons under a common climate, the resulting SCOP was a model-derived estimate for conditions not necessarily experienced by an installation, rather than its achieved efficiency over a specific observation period. Moreover, the approach required sufficient

observations for system-specific model fitting, did not account for domestic hot water (DHW) in the SCOP evaluation, and did not establish product-specific performance expectations. The complementary challenge is therefore to estimate the efficiency expected from a given heat pump model under its observed operating conditions, rather than translating observed performance to standardized conditions.

Here, we estimate expected field performance by combining model-specific laboratory coefficient of performance (COP) data, thermodynamic temperature-lift relationships, and an empirically calibrated field correction. This provides a product-specific expectation under observed operating conditions, accounting for differences in laboratory-rated efficiency between heat pump models. Expected SPF is calculated separately for central heating (CH) and DHW and combined according to their thermal-energy contributions. Comparing achieved and expected SPF defines target achievement, allowing performance evaluation from monthly observations without fitting a model to each installation. The resulting reference accounts for observed operating conditions but does not determine whether these conditions could be improved, for example by reducing heating supply temperatures. We develop and validate the approach using field data from 176 heat pumps and apply it without recalibration to 10,087 installations. Systems achieving less than 80% of their expected performance had 2.3 times the incidence of recorded service contacts, whereas an absolute SPF threshold of 3.0 showed little difference between groups. These findings indicate that context-adjusted performance evaluation can help prioritize installations for further investigation using routinely available operational data.

2 Results

2.1 Estimating expected performance under field conditions

The aim of this study is to estimate the seasonal performance factor that can reasonably be expected from an individual heat pump under observed operating conditions. The approach considers outdoor and supply temperatures, which determine the temperature lift; model-specific laboratory COP values, which characterize product performance; and the contributions of CH and DHW, which account for differences in operating temperature requirements. We develop the corresponding estimation procedure and compare alternative formulations for translating laboratory-based performance to field conditions.

The influence of temperature lift on heat pump efficiency is well established, and previous field studies have evaluated heat pump performance relative to the thermodynamic Carnot limit using a Carnot efficiency factor, reaching around 0.39 [5, 6]. This factor reflects both the efficiency of the heat pump itself and the effects of its operation under field conditions. We build upon this approach and distinguish these two contributions using two separate factors: a product-specific laboratory factor derived from standardized coefficient of performance (COP) measurements, describing how efficiently the heat pump approaches the Carnot limit under test conditions, and an empirical field factor capturing additional efficiency losses during installed operation, including effects associated with transient operation, cycling, control strategies, and installation practices. This distinction allows the efficiency of the heat pump itself, as determined under laboratory conditions, to be evaluated separately from the effects of real-world installation and operation. Figure 1 illustrates the progression from thermodynamic potential to expected and achieved field performance.

Performance is evaluated at monthly resolution to balance short-term variability with timely assessment. Daily efficiency fluctuates with weather and operating conditions, whereas annual evaluation delays feedback. Previous research has demonstrated the feasibility of estimating seasonal performance from shorter monitoring periods, recommending at least 10 days of observations [12]. Monthly performance indicators are also well established in field-monitoring studies [10, 14].

2.1.1 Complementary field datasets for model estimation

The approach was developed and evaluated using three complementary reference datasets comprising 176 residential air-to-water heat pumps: five systems from a supervised Swiss field study with calibrated measurement equipment, 104 from HeatpumpMonitor.org with independently metered heat and electricity, and 67 quality-screened installations monitored through manufacturer telemetry. The manufacturer subset was selected using installation-quality indicators independent of measured efficiency, while the open-source systems were screened for comparable metering boundaries and data completeness. For all systems, model-specific laboratory COP values were retrieved from the KEYMARK database [15], and the estimation was conducted separately for CH and DHW at monthly resolution. Together, these sources

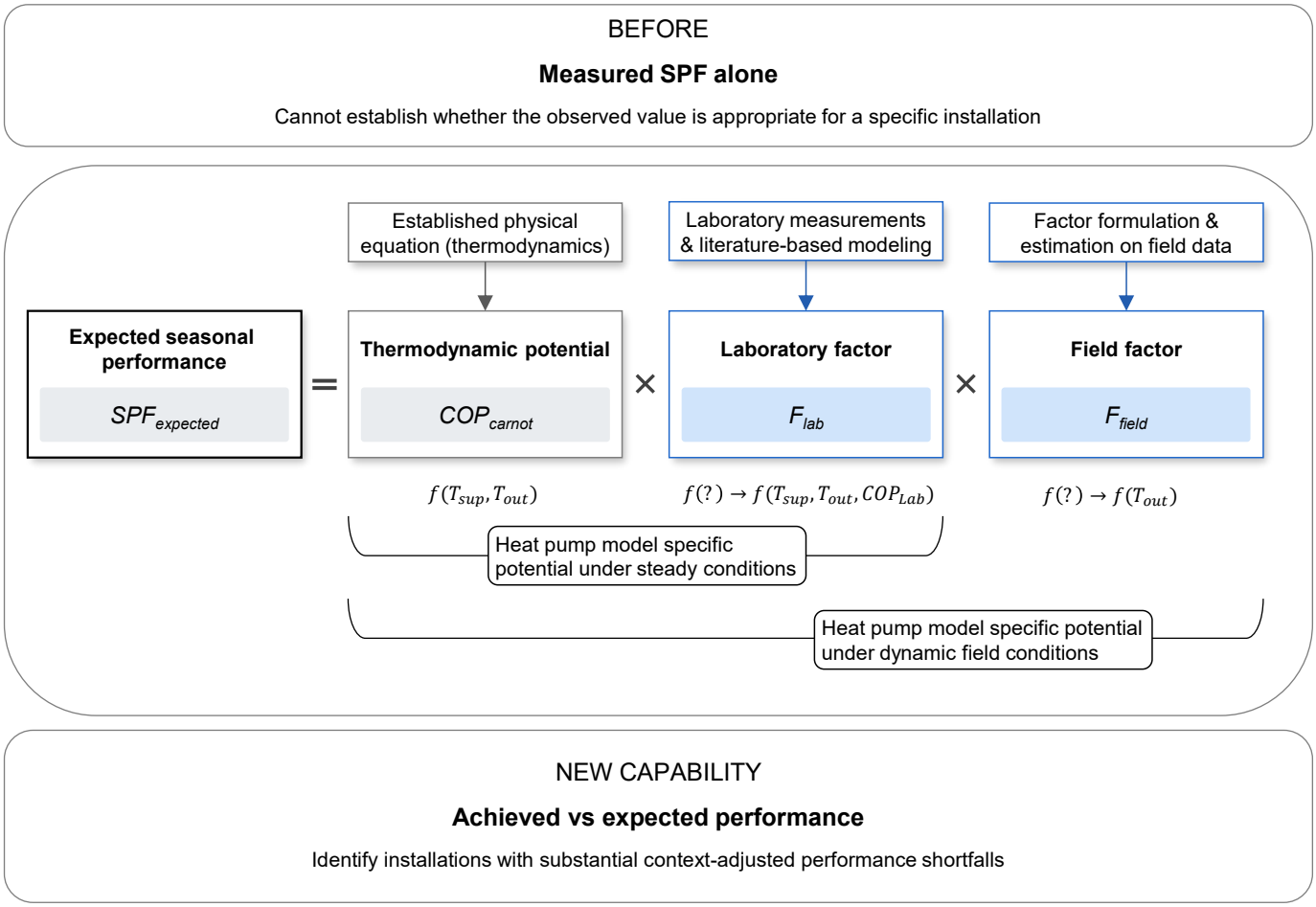


Figure 1: Conceptual framework for context-specific heat pump performance expectation and evaluation. Thermodynamic temperature-lift dependence provides the physical reference, while model-specific laboratory COP data characterize the performance potential of the installed heat pump model. An empirical field correction translates this laboratory-based potential to expected field performance under application-specific operating conditions. These conditions may be anticipated during system design or observed after installation. In the present study, the framework is developed and evaluated using observed field conditions. After installation, comparison of achieved and expected SPF defines target achievement and enables context-adjusted performance screening.

provide different measurement approaches and operating contexts for developing and evaluating the model.

The datasets represent substantially different operating conditions (Fig. A1). The open-source systems experienced milder outdoor temperatures during the heating season than the Swiss and German installations (8.1 versus approximately 5.2 °C). More pronounced differences occurred in CH supply temperatures evaluated at an outdoor temperature of -10 °C, averaging 63.8 °C in the German dataset compared with approximately 38 °C in the Swiss and open-source datasets. Consistent with these differences in temperature lift, mean achieved total SPF ranged from 2.88 in the German dataset to 3.98 in the open-source dataset, with 3.42 in the Swiss study. These contrasts illustrate why differences in measured efficiency cannot be interpreted independently of operating conditions.

The datasets also differ in thermal energy demand and operating-mode contributions. Mean annual thermal output ranged from 9.9 MWh in the open-source dataset to 19.4 MWh in the small Swiss sample, while the average DHW contribution was relatively similar across datasets (17-23%) but varied considerably between individual installations. This contribution makes DHW relevant to the evaluation of total system efficiency, rather than treating space-heating performance as representative of the entire installation. These characteristics provide additional variation in heating requirements and operating-mode composition. The resulting heterogeneity forms the basis for estimating the field correction and evaluating its transferability across distinct operating conditions and monitoring systems, rather than relying on a single type of installation. Detailed dataset characteristics and selection criteria are provided in Methods.

2.1.2 Selection and evaluation of the field-correction model

We compared nine candidate formulations to determine how product-specific laboratory performance can best be translated to field conditions. The formulations explored additive and multiplicative corrections, different dependencies on operating conditions, including outdoor and supply temperatures, temperature variability, and system utilization, as well as corrections applied either directly to predicted performance or to Carnot-normalized efficiency. A multiplicative field correction depending on outdoor temperature provided the best cross-validated performance for both CH and DHW (Appendix Table A2). Expected performance was therefore calculated as

$$\mathrm{SPF}_{\mathrm{expected},m} = \mathrm{COP}_{\mathrm{Carnot}}(T_{\mathrm{out}}, T_{\mathrm{supply}}) \cdot F_{\mathrm{lab},m}(T_{\mathrm{out}}, T_{\mathrm{supply}}, \mathrm{COP}_{\mathrm{lab}}) \cdot F_{\mathrm{field},m}(T_{\mathrm{out}})\,. \tag{1}$$

Here, *m* denotes the operating mode (CH or DHW), while T_{out} and T_{supply} denote mean outdoor and supply temperatures, respectively. COP_{Carnot} represents the thermodynamic efficiency limit, F_{lab} accounts for product-specific performance based on laboratory COP measurements (COP_{lab}), and F_{field} corrects for systematic differences between laboratory-based expectations and observed field performance. Both correction factors are mode-specific.

The selected approach achieved cross-validated median absolute percentage error (MdAPE) values of 6.63% for CH, 8.39% for DHW, and 7.03% for total performance, with corresponding R^2 values of 0.76, 0.43, and 0.65 (Table 1). We further examined whether deviations from expected performance varied primarily between installations or across observation periods. Between-system variability in target achievement was greater for CH and DHW, whereas within-system variability was greater for total performance. The selected field correction was subsequently re-estimated using the complete reference dataset for fleet-scale application. Detailed validation statistics and final model parameters are provided in Appendix Tables A2 and A4.

Table 1: Cross-validated prediction accuracy and variation in target achievement. Target achievement is reported as mean ± standard deviation. Between-system and within-system standard deviations describe variation across installations and observation periods, respectively, in percentage points (pp). N denotes monthly observations (system-months), and n denotes individual heat pump systems.

Mode	MAE	MdAPE (%)	RMSE	R^2	TA μ ± σ	Between SD (pp)	Within SD (pp)	N	n
CH	0.33	6.63	0.45	0.76	100.00 ± 11.09	9.77	7.74	969	169
DHW	0.34	8.39	0.49	0.43	100.02 ± 16.04	14.45	9.95	1,621	175
Total	0.34	7.03	0.47	0.65	99.88 ± 13.74	9.85	10.72	1,654	176

Figure 2 illustrates how the resulting expected performance varies with outdoor and supply temperatures for CH and DHW. As expected from the thermodynamic temperature lift, efficiency decreases with lower outdoor and higher supply temperatures. The figure uses representative laboratory COP values for illustration, whereas expected performance is calculated using product-specific laboratory data for each installation. As reference input, we use the fleet-average laboratory COP values at the operating points $A{-}7/W35$, $A2/W35$, $A7/W35$, $A{-}7/W55$, $A2/W55$, and $A7/W55$, which are 2.60, 4.74, 6.15, 2.11, 3.54, and 4.65, respectively.

Magnitude of laboratory and field correction factors. The laboratory factor increased with outdoor temperature, while the empirical field correction decreased, partly offsetting this increase. At an outdoor temperature of 7 °C, close to the reference datasets' mean, the median laboratory factor was approximately 0.55 for CH at 35 °C supply temperature and 0.65 for DHW at 55 °C. The corresponding field factors were 0.63 and 0.56. Together, these factors indicate that approximately 35–36% of the theoretical Carnot efficiency would be expected under these illustrative conditions. For comparison, Fraunhofer ISE reported Carnot efficiencies of 0.32–0.44 (mean 0.39), while HeatpumpMonitor.org reports approximately 0.45–0.55 [6, 16]. These values are not directly comparable because each approach uses slightly different temperatures to calculate the theoretical maximum. Fraunhofer uses the temperature entering the heat pump on the source side and the average heating-water temperature, while HeatpumpMonitor.org assumes refrigerant temperatures 2 K above the supply-water temperature and 6 K below the outdoor-air temperature. Our approach uses the measured outdoor-air and supply-water temperatures directly, as these were available across all datasets. Therefore, differences in the reported fractions do not necessarily indicate differences in actual heat pump performance.

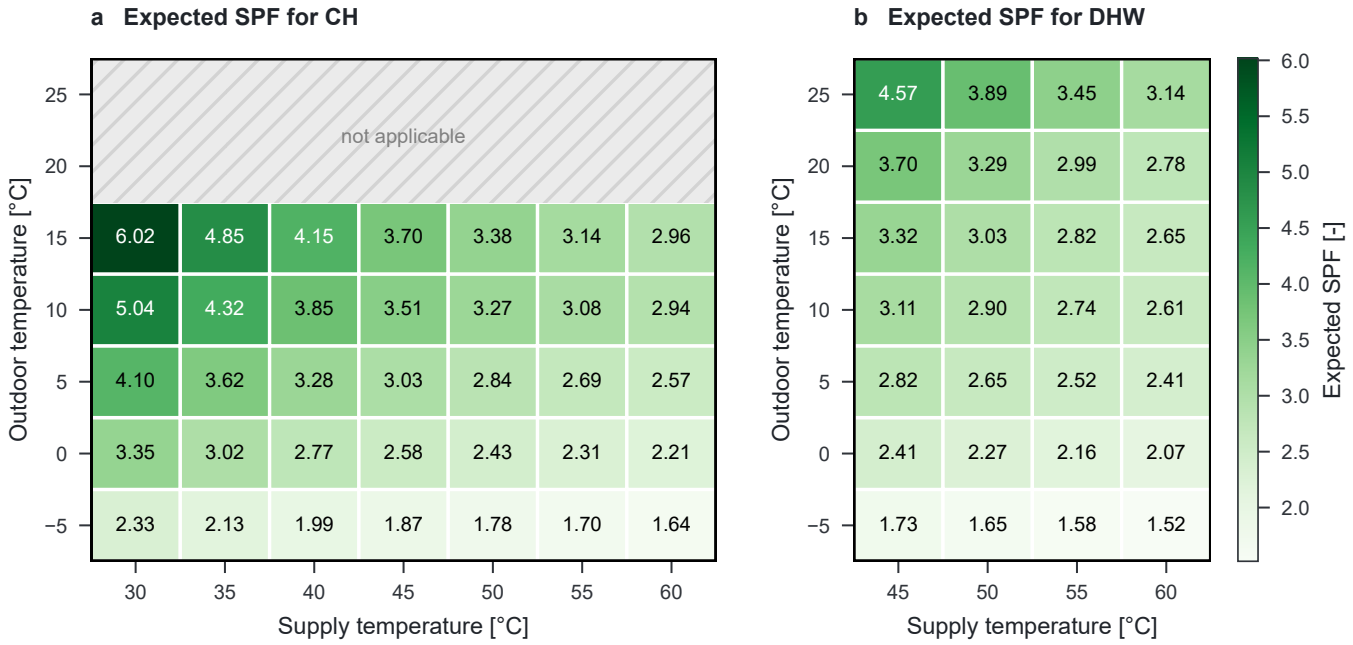


Figure 2: Expected heat pump performance under representative operating conditions for **a** central heating and **b** domestic hot water. The values are calculated with the proposed method, using exemplary laboratory COP values at the standard operating points $A{-}7/W35$, $A2/W35$, $A7/W35$, $A{-}7/W55$, $A2/W55$, and $A7/W55$. The heatmaps illustrate how the expected performance factor decreases with lower outdoor temperature and higher supply temperature. Grey hatched areas indicate operating conditions that are not applicable for the respective mode.

Accounting for operating-temperature effects Across the pooled system-grouped out-of-fold predictions, mean target achievement remained close to 100% for CH, DHW, and total performance (99.88% for the combined total), without a pronounced seasonal pattern. Furthermore, outdoor and supply temperatures explained little of the remaining variation in target achievement (R^2 at most 0.032), indicating that the model captures most of their systematic linear influence. Detailed distributions, monthly trends, and residual analyses are provided in the Appendix A2.2 and Table A3.

Transferability across monitoring datasets To examine whether the approach generalizes beyond the data used for calibration, we performed leave-one-dataset-out assessment. In each iteration, the field correction was estimated using two reference datasets and evaluated on the third, previously withheld dataset. The resulting MdAPE ranged from 5.33% to 7.85% for CH and from 6.73% to 13.11% for DHW. Mean target achievement ranged from 94.93% to 106.66% across the withheld datasets. These findings support transferability across heterogeneous monitoring sources, although remaining dataset-specific deviations indicate that calibration is not fully independent of the monitored population (Appendix Table A5).

Robustness across observation periods Heat pump performance is commonly evaluated over monthly or annual periods in field-monitoring studies [10, 14], while previous research has also explored estimating seasonal performance from shorter monitoring periods [12]. Since our model was calibrated using monthly observations, we examined whether the same field-correction parameters remain applicable at annual resolution without re-estimation. Applied to annually aggregated reference data, mean target achievement remained close to 100% for CH (102.14%), DHW (100.98%), and combined performance (102.12%). For combined performance, the MdAPE was 5.04%, with an R^2 of 0.756. In the larger manufacturer fleet, annual predictions also retained moderate absolute errors but explained considerably less variation between installations, with R^2 values of 0.149 for CH, 0.035 for DHW, and 0.205 for total performance. These findings indicate that the monthly-derived correction can also be applied to annual performance assessment, although its ability to explain differences between installations varies across datasets. Its reliability for observation periods shorter than one month remains untested. Detailed residual analyses are provided in Appendix Section A2.

2.2 Fleet-scale performance assessment and service demand

To evaluate the applicability of the proposed approach at scale, we applied the previously estimated model without further recalibration to operational data from 10,087 residential air-to-water heat pumps. The systems originate from a single manufacturer, are installed predominantly in Germany, and were monitored between April 2025 and April 2026. Unlike the Swiss and open-source monitoring datasets, the manufacturer telemetry provides estimates of electricity consumption and thermal energy output derived from sensors integrated into the heat pumps, which is common practice in most modern heat pumps [17]. The fleet includes a broad range of operating conditions and installations, allowing us to examine the distribution of achieved efficiency and context-adjusted target achievement in practice.

In addition to the telemetry data, anonymized service records covering November 2025 to February 2026 were linked to the monitored heat pumps. The resulting analysis included 9,604 systems, of which 1,017 (10.6%) had at least one recorded service contact. These records include customer inquiries, technical support requests, and fault-related service contacts, and therefore do not necessarily indicate technical faults. We subsequently investigated whether context-adjusted underperformance was associated with increased service demand and operational anomalies, and compared these associations with those obtained using absolute SPF thresholds. Further details are provided in Methods.

2.2.1 Identifying potentially underperforming installations

Absolute and context-adjusted efficiency capture different aspects of heat pump performance. A low achieved SPF may reflect demanding operating conditions rather than poor system operation [13], whereas a substantial shortfall from expected performance can also occur at relatively high absolute efficiency. We therefore considered both criteria when identifying potentially underperforming installations. An absolute SPF threshold of 3.0 was selected given the predominantly German fleet, its use in previous literature, and earlier German building-efficiency funding requirements, which specified a calculated annual performance factor (Jahresarbeitszahl (JAZ)) of at least 3.0 according to VDI 4650 Blatt 1 [18]. Although this design-stage criterion is not directly equivalent to measured SPF, it provides a policy-relevant reference for combined performance. The target-achievement threshold was set at 80%, representing a performance shortfall of more than 20% relative to expectation, substantially exceeding the median prediction errors reported above. Both thresholds serve as screening criteria rather than established technical fault boundaries. Their influence on the classification was therefore evaluated across alternative threshold combinations. Full analysis is available in Appendix Section A2.6.

At the selected thresholds, 7.8% of CH systems, 4.7% of DHW systems, and 5.1% of systems based on total performance had both an achieved SPF below 3.0 and target achievement below 80%. In comparison, an absolute SPF below 3.0 alone identified 46.6%, 92.8%, and 60.5%, respectively. The joint classification thus distinguishes installations with low absolute efficiency from those additionally performing substantially below their context-specific expectation. The proportion of systems identified as potentially underperforming depends on the selected target-achievement threshold. Keeping the absolute SPF threshold fixed at 3.0, lowering the target-achievement threshold from 80% to 70% reduced the proportion of identified systems from 5.1% to 0.3%, whereas increasing it to 90% raised this proportion to 26.0% (Fig. A2). These findings demonstrate the additional selectivity of context-adjusted screening while highlighting the importance of threshold choice.

2.2.2 Performance groups identified by context-adjusted screening

The combination of achieved and expected SPF distinguishes three forms of context-adjusted underperformance (Fig. 3). All three groups have target achievement below 80%. Zone I comprises systems with achieved SPF below 3.0 despite an expectation of at least 3.0. In Zone II, both achieved and expected SPF are below 3.0, indicating demanding operating conditions combined with an additional performance shortfall. Zone III comprises systems with achieved SPF of at least 3.0 that nevertheless fall substantially below their higher expectation.

For total performance, 3.0% of all evaluated systems belonged to Zone I, 2.2% to Zone II, and fewer than 0.1% to Zone III. Thus, 5.1% combined low absolute efficiency with substantial context-adjusted underperformance. The distributions differed between operating modes: 6.0% of systems evaluated for CH belonged to Zone I, whereas 4.6% of systems evaluated for DHW belonged to Zone II. These findings illustrate how similar absolute efficiency levels can reflect different operating contexts and deviations from expected performance.

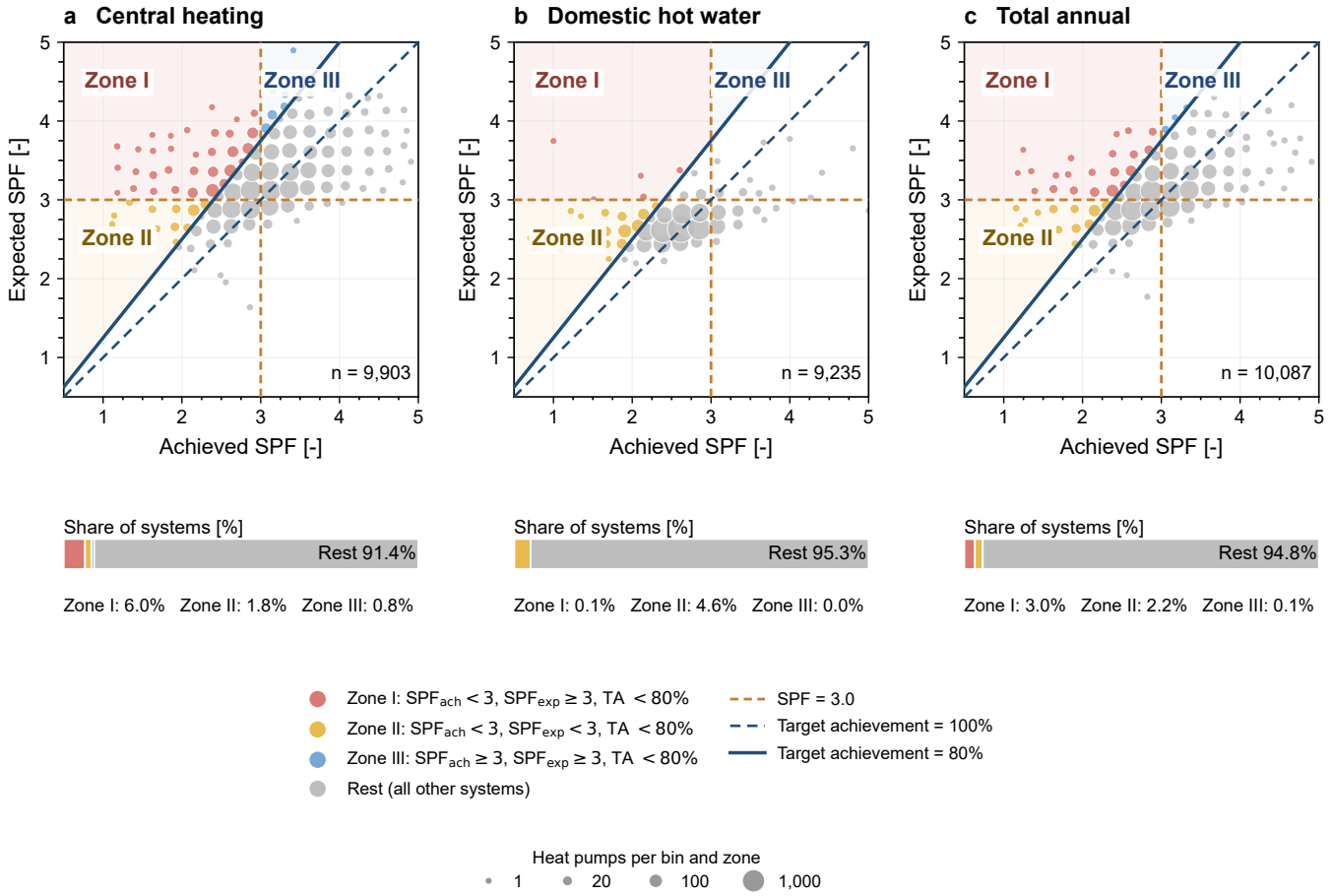


Figure 3: Large-scale application of the target-achievement metric to 10,087 air-to-water heat pumps. **a**, central heating; **b**, domestic hot water; and **c**, combined performance. Upper panels compare achieved and expected SPF, with marker size indicating the number of systems per bin and colour distinguishing performance groups. Dashed lines indicate SPF = 3.0 and 100% target achievement, while the solid blue line marks 80% target achievement. Lower panels show the proportion of systems in each group, as defined in the legend.

2.2.3 Operational characteristics of underperforming systems

The identified performance groups differed in thermal energy characteristics, operational behaviour, heat pump characteristics, and operating conditions (Fig. 4). Zone I (n=300) comprised relatively small heat pumps, with 16% lower nominal capacity, 9% lower total thermal energy output, and 12% lower CH supply temperatures than the fleet average. Despite these comparatively favourable temperature requirements, compressor-alarm duration was 7.27 times and auxiliary-electric-heater energy 7.19 times the fleet average. These substantial deviations indicate that the low achieved efficiency of Zone I systems is associated with operational irregularities rather than particularly demanding temperature requirements.

Zone II (n=218) showed a different operating profile, combining higher temperature requirements with pronounced operational deviations. Compared with the fleet average, these systems had a 57% higher DHW share, 26% greater DHW energy output, and 9% higher CH supply temperatures, while CH energy output was 26% lower. Compressor starts were 33% higher, compressor-alarm duration was 7.03 times the fleet average, and auxiliary-electric-heater energy was 8.87 times higher. Thus, Zones I and II exhibited similarly elevated operational indicators despite different temperature requirements and heating-demand profiles.

Zone III (n=9) comprised systems with relatively low heating demand, DHW contributions, and CH supply temperatures. Unlike Zones I and II, these systems showed below-average compressor-alarm duration and auxiliary-electric-heater energy, despite achieving substantially less than their expected SPF. However, the small sample limits the interpretation of these differences. Laboratory-weighted performance factors remained relatively close to the fleet average across all groups, suggesting that the observed

differences are not primarily associated with the laboratory-rated product characteristics represented by these factors. Overall, the results reveal distinct operating profiles among underperforming systems, although the observed associations do not establish the causes of individual performance shortfalls.

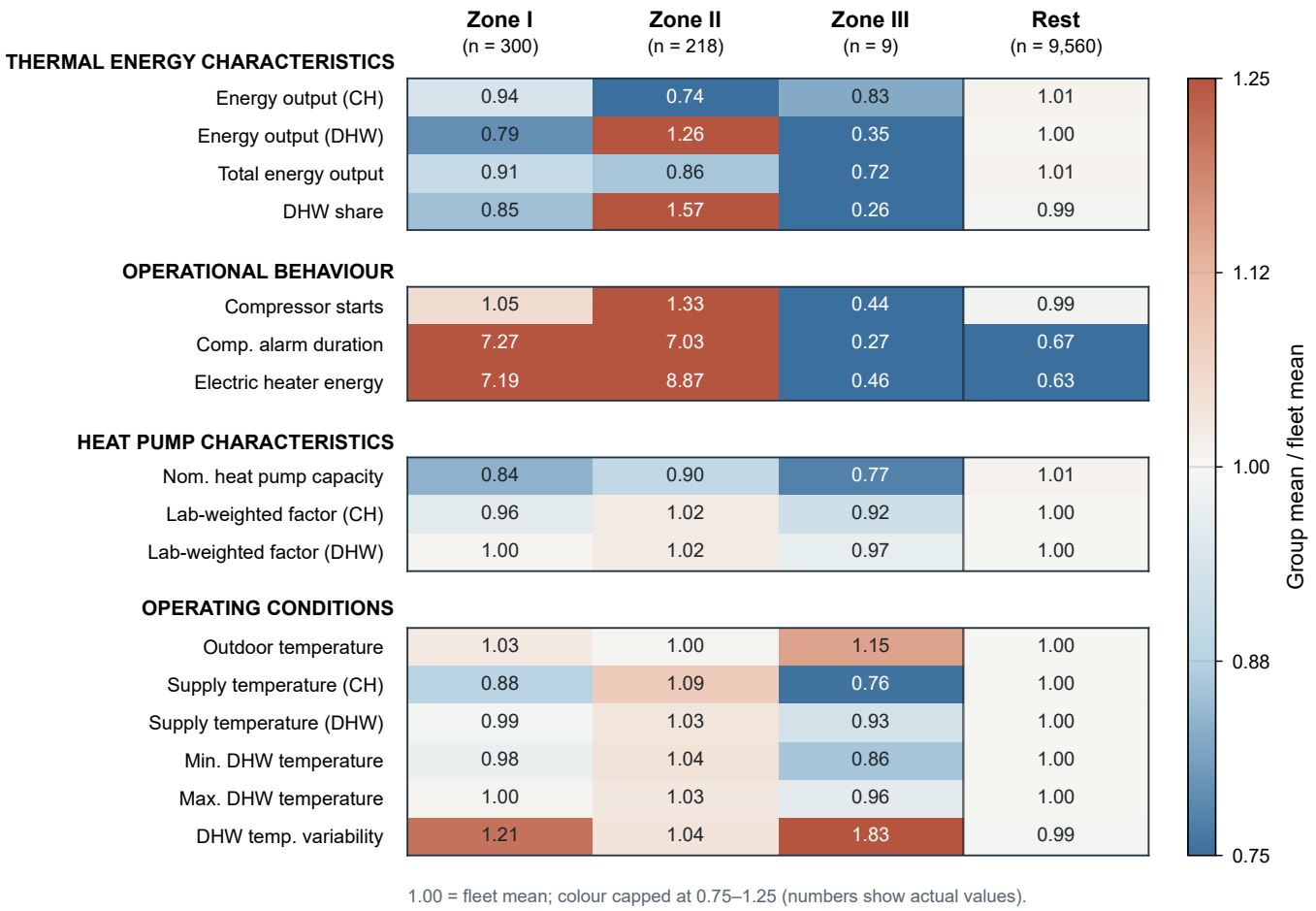


Figure 4: Characteristics of heat pumps across performance groups. Cells show the ratio of the group mean to the fleet mean for each variable, grouped into thermal energy characteristics, operational behaviour, heat pump characteristics, and operating conditions. Values above and below 1 indicate higher and lower group means relative to the fleet, respectively. The colour scale is capped at 0.75–1.25 for visual clarity, while cell labels report the actual ratios, including values outside this range. Sample sizes are indicated above each group. All three underperformance zones have target achievement below 80%: Zone I comprises systems with achieved SPF below 3.0 and expected SPF of at least 3.0; Zone II comprises systems with both achieved and expected SPF below 3.0; and Zone III comprises systems with both achieved and expected SPF of at least 3.0. Rest comprises all remaining systems.

2.2.4 Association with recorded service demand

To investigate whether performance deviations precede increased service demand, we compared monthly SPF and target achievement with recorded service contacts in the subsequent month (Fig. 5). Unlike the full-period fleet assessment in Fig. 3, this analysis captures performance shortly before potential service activity, meaning that system classifications may differ between the two observation periods. Among the 9,604 systems with linked service records, 22.2% of those achieving less than 80% of expected performance had at least one subsequent service contact, compared with 9.6% of the remaining systems. This corresponds to a 2.3-fold higher service-contact incidence.

By contrast, an absolute SPF threshold of 3.0 showed little difference, with service-contact incidence of approximately 10.6% in both groups. The mean number of recorded contacts also differed more strongly between target-achievement groups (0.39 versus 0.15 contacts per system) than between absolute SPF groups (0.18 versus 0.15). These findings indicate that context-adjusted performance deviations are more

strongly associated with subsequent service demand than absolute efficiency alone, although service contacts do not necessarily indicate technical faults and the observed associations do not establish causality.

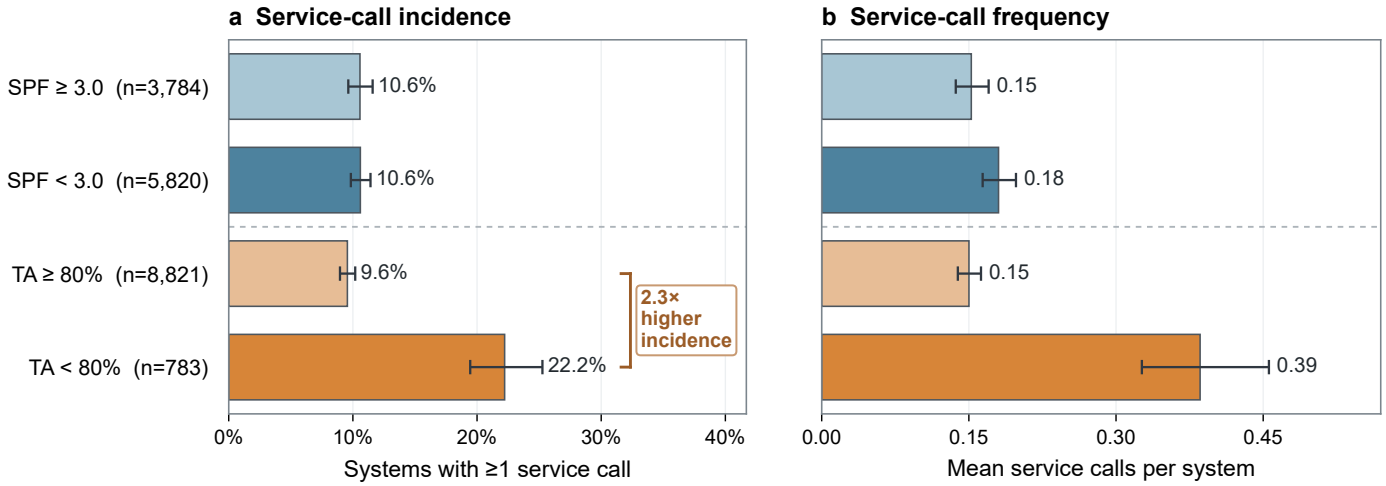


Figure 5: Association of target achievement and absolute SPF with recorded service activity. **a**, Fraction of systems with at least one recorded service contact during the service-reporting period; error bars indicate 95% Wilson score confidence intervals. **b**, Mean number of recorded service contacts per system; error bars indicate 95% confidence intervals estimated using negative-binomial models. Both panels use the same system-level analysis cohort and thresholds of TA=80% and SPF=3.

3 Discussion

3.1 From absolute efficiency to context-adjusted performance

Field studies have reported substantial differences between laboratory-rated and installed heat pump efficiency [17, 19, 20]. However, these differences do not necessarily indicate deficiencies in the heat pump itself. Miara et al. found average deviations below 10% between laboratory and field COP when measurements were restricted to comparable quasi-steady operating conditions [10]. Such comparisons assess product performance at selected operating points but exclude much of the dynamic operation relevant to seasonal efficiency. For example, Bernal et al. identified quasi-steady conditions during only approximately 17% of monitored space-heating operation [21]. Cycling, defrosting, installation characteristics, and system controls can further affect performance over longer periods [19]. The relevant question is therefore not whether a heat pump reproduces its rated COP under selected conditions, but whether it achieves reasonable efficiency throughout its actual operation.

Established methods such as VDI 4650 estimate annual efficiency using product performance data and standardized design assumptions, but do not provide a reference for specific monitoring periods. Carnot-based assessments, already used by Fraunhofer ISE and HeatpumpMonitor.org, account for operating temperatures [6, 16]. However, translating the theoretical Carnot limit into expected field efficiency requires both a consistent definition of the temperatures and an appropriate efficiency factor. Our findings show that this factor depends on product characteristics and operating conditions, making a single constant value, such as 0.35, insufficient for evaluating individual installations.

Previous field-based normalization to EN 14825 reference conditions relied on system-specific heating-curve and COP models to estimate space-heating SCOP [13]. Although useful for standardized comparisons, these estimates could require extrapolation to conditions not observed by individual installations and did not directly represent the ratio of delivered heat to consumed electricity over the monitoring period. The need for sufficient observations and suitable model fits also restricted the earlier efficiency evaluation to 708 of 1,023 installations (69.2%). The present approach instead evaluates achieved SPF under observed operating conditions against an independently estimated, product-specific expectation. By eliminating system-specific model fitting, it allows assessment from monthly observations without requiring the operating-condition variation needed to fit individual models. Incorporating laboratory COP characteristics also means that heat pump models with higher certified efficiency are assigned higher expectations

under comparable conditions. Furthermore, the evaluation extends beyond CH to include DHW, which contributed approximately 17–23% of useful heat output on average across the reference datasets. Together, these changes provide a more directly interpretable measure of observed performance relative to expectation, while reducing the dependence on extensive system-specific monitoring histories.

The results suggest that such a reference can be estimated with useful accuracy from routinely available operating data, despite differences between installations and monitoring sources. Cross-validation yielded median absolute percentage errors of 6.6% for CH and 8.4% for DHW, while leave-one-dataset-out assessment supported transferability across the available datasets. The greater uncertainty for DHW also highlights the limits of representing complex operation through monthly averages. Intermittent operation, storage losses, draw-off patterns, and control strategies may introduce variation not captured by the available temperatures and laboratory characteristics. Similar influences of cycling, defrosting, and auxiliary heating on field efficiency have been identified in previous studies [17, 22–24]. Thus, the approach provides a practical expectation for performance screening without requiring a detailed representation of every operational process.

3.2 Operational profiles and service relevance

Combining target achievement with absolute SPF provides a more differentiated basis for interpreting low heat pump efficiency. Zone I systems, which operate under comparatively favourable temperature conditions but exhibit elevated operational indicators, appear particularly relevant for diagnostic follow-up. Zone II systems instead combine demanding heating requirements with additional performance shortfalls, suggesting that both system design and operational behaviour warrant investigation. Previous studies have identified elevated supply temperatures, auxiliary heating, compressor cycling, and control settings as important influences on seasonal efficiency [17, 19, 23, 25, 26]. The identified profiles can therefore help direct further investigation, although the observed indicators cannot establish whether performance losses arise from faults, system configuration, or operating conditions not captured by the model.

Zone III illustrates a different limitation of absolute efficiency thresholds: systems may achieve seemingly satisfactory efficiency while still performing substantially below their product-specific expectation. The absence of pronounced alarm activity or auxiliary-heater use suggests that such deviations need not coincide with obvious operational warning signs. However, the small number of systems in this group limits further interpretation. Overall, the classification provides a basis for distinguishing different diagnostic priorities rather than identifying specific fault types.

The association between low target achievement and subsequent service contacts further supports the practical relevance of context-adjusted screening. Previous research has shown that energy-monitoring data can help preselect heat pumps for professional inspections [4]. By incorporating product-specific performance and observed operating conditions, the present approach provides a more differentiated reference for identifying installations that may warrant attention. However, service contacts do not necessarily indicate technical faults, and the observed association does not establish causality or demonstrate that the identified performance shortfalls are correctable.

3.3 Implications for homeowners, industry, and policy

For homeowners, the proposed approach could make performance feedback more interpretable and actionable. Reporting only an achieved SPF leaves users without a reference for deciding whether the value is appropriate for their home and heat pump product. This information gap matters because household satisfaction with heat pumps is influenced by operating and maintenance costs, while a survey of heat pump users found that 68% desired feedback on fuel and cost savings and 58% on system efficiency [27]. Target achievement could help homeowners distinguish between low efficiency that is broadly consistent with observed operating conditions and performance that falls substantially below expectation. In the former case, attention could turn towards reducing temperature requirements through improved system settings or building insulation, whereas the latter may warrant investigation of operational inefficiencies or technical issues. However, the approach does not establish whether operating conditions are optimal or identify specific faults, and such feedback should guide users towards appropriate professional support rather than encourage unsupervised changes [28, 29].

For manufacturers and service providers, the principal implication is more efficient allocation of diagnostic resources. Connected heat pumps already generate the temperature, operational-state, and energy data required for remote monitoring, yet many existing services remain focused on displaying consumption or reporting explicit alarms [30, 31]. Target achievement adds a fleet-level screening layer before detailed fault diagnosis, helping prioritize installations and guide subsequent investigations.

Systems in Zone I could be prioritized because low efficiency is not readily explained by the evaluated context, whereas Zone II systems may require a broader assessment of temperature requirements, DHW demand, control, and system design. Combining context-adjusted benchmarking with complementary assessments of heating-curve settings and temperature requirements could provide a more comprehensive basis for identifying optimization potential and supporting installation quality assurance. Zone III systems may warrant lower-priority optimization review despite acceptable absolute efficiency. This hierarchy could reduce unnecessary on-site visits while increasing the likelihood that limited service capacity is directed towards systems with potentially recoverable efficiency losses. Such prioritization is increasingly relevant because heat pump deployment is already constrained by shortages of qualified installation and service personnel [3].

For installers, repeated assessment during commissioning and early operation could provide a feedback loop that is largely absent from one-off design and installation processes. Installers are influential intermediaries in technology selection, system sizing, configuration, and user guidance, but they face rapidly changing technologies and substantial uncertainty when estimating building demand and realized energy use [32–34]. Unlike conventional fault alarms, target achievement can reveal performance shortfalls even when no explicit technical fault is recorded, helping installers assess whether a system operates efficiently under its actual conditions. This would complement established field-monitoring and quality-assurance initiatives, which have identified substantial optimization potential through improved system sizing, control settings, and commissioning practices [5, 6, 35]. Aggregated results across an installer's portfolio could also support learning by showing whether particular design or commissioning practices are systematically associated with lower target achievement. The proposed approach cannot attribute deficiencies to installers, because building conditions, product behaviour, occupants, and subsequent control changes also influence performance. Its value lies in identifying cases and patterns that merit closer examination.

For policymakers, context-adjusted performance could complement rather than replace standardized product ratings. Energy labels influence technology choices and provide comparable information at the point of purchase [36], while standardized laboratory conditions remain necessary for fair product comparison. However, interpreting standardized ratings as achievable field performance may create unrealistic expectations [17, 19]. The automotive industry illustrates the broader consequences of such discrepancies, where substantial gaps between official fuel-consumption ratings and real-world performance have prompted public criticism of manufacturers and testing procedures [37]. Similar discrepancies in heat pump performance could undermine consumer confidence even when products operate as expected under their actual conditions. The proposed approach could help bridge this gap by translating product-specific laboratory performance into realistic expectations for individual installations. Aggregated target achievement distributions could also provide an outcome-oriented indicator for installation-quality programs, subsidy evaluations, or large-scale monitoring initiatives while avoiding the assumption that all homes should achieve the same absolute SPF. Such use could improve the evaluation of public expenditure on heat pump deployment [38, 39]. The present thresholds should not yet be adopted as regulatory fault boundaries, because they are empirical screening choices derived predominantly from one manufacturer's fleet and require independent validation.

Although the approach was evaluated retrospectively, its structure also suggests a possible connection between design-stage expectations and post-installation assessment. Established procedures such as VDI 4650 estimate prospective seasonal performance from assumed climate and system characteristics. If anticipated outdoor and supply temperatures and the expected contributions of CH and DHW are available during design, the present formulation could similarly provide an application-specific expected SPF. Observed conditions could then update this expectation after commissioning, and target achievement could quantify whether realized performance remains consistent with the updated reference. A continuous reference across design, commissioning, and operation could improve operating-cost estimates, which are central to heat pump total-cost-of-ownership assessments [40, 41]. This prospective application was not validated here and should be treated as a research direction rather than as a replacement for established design-stage calculations.

The proposed approach may also be applicable beyond manufacturer platforms. Smart-meter data have demonstrated potential for heat pump identification, preselection, and performance-related feedback when direct telemetry is unavailable [4, 22, 42, 43]. However, estimating target achievement requires information on useful heat output, operating temperatures, operating-mode contributions, and product-specific laboratory performance that conventional electricity meters do not provide directly. Manufacturer-independent application would therefore require additional sensing, standardized data interfaces, or validated methods for estimating these inputs. This requirement highlights the value of interoperable and privacy-preserving access to operational heat pump data.

3.4 Limitations and future research

Several limitations constrain interpretation and generalization. First, the empirical field correction was derived from curated field datasets that do not necessarily represent fault-free or optimally operated systems. Expected SPF therefore reflects typical performance under observed conditions rather than a technical optimum. Suboptimal supply temperatures, weather compensation, cycling, or system sizing may reduce both expected and achieved performance, such that high target achievement can coexist with remaining optimization potential [6, 7, 35].

Second, the reference datasets differ in sample size, measurement equipment, temporal resolution, processing, and system boundaries, contributing to dataset-specific bias. Leave-one-dataset-out assessment supports transferability across the available sources, but further independent validation is required, particularly for sparsely represented extreme operating conditions.

Third, the laboratory translation relies on fixed modelling assumptions necessitated by sparse product data, including interpolation and bounded extrapolation of Carnot-normalized efficiency, within-period temperature smoothing, a common heating-balance temperature, and fixed outdoor-temperature bins. These choices were specified independently of achieved field performance but were not validated individually; the out-of-sample evaluation therefore assesses the formulation as a whole rather than each assumption separately.

Fourth, the large-scale application represents predominantly recent heat pumps from one manufacturer and geographic region. Generalization to other manufacturers, climates, refrigerants, building types, distribution systems, and control concepts remains uncertain. Manufacturer energy quantities are partly controller-derived rather than independently metered, and differences in measurement boundaries may introduce additional uncertainty [19]. Monthly aggregation further limits sensitivity to short-duration phenomena such as defrosting, auxiliary-heater operation, and transient faults.

Fifth, relevant installation characteristics such as hydraulic configuration, building heat demand, storage operation, emitter characteristics, occupant behaviour, and control settings were unavailable or inconsistent. Low target achievement therefore indicates a performance deviation but cannot identify a specific underlying cause.

Finally, service records covered only November 2025 to February 2026 and did not distinguish between technical faults, commissioning issues, remote support, or on-site interventions. The observed service associations are therefore descriptive rather than causal. Likewise, the thresholds of SPF=3.0 and target achievement of 80% are screening choices rather than fault boundaries. Future work should validate the method and threshold selection using additional manufacturers, independently diagnosed faults, interventions, realized energy savings, and longer observation periods.

4 Conclusion

This study introduces a method for estimating the SPF expected from residential heat pumps under their observed operating conditions. By combining product-specific laboratory data with thermodynamic relationships and an empirical field correction, the method provides a reference for evaluating individual installations without system-specific model training. Cross-validated median prediction errors were 6.6% for CH and 8.4% for DHW, with applicability demonstrated at monthly and annual resolution.

Application to 10,087 heat pumps demonstrated the practical value of evaluating efficiency relative to expectation. Systems achieving less than 80% of expected performance in the month preceding service activity had a 2.3-fold higher incidence of subsequent service contacts, whereas an absolute SPF threshold of 3.0 showed little discrimination. Furthermore, systems with unexpectedly low absolute efficiency (Zone I) exhibited 7.27 times greater compressor-alarm duration and 7.19 times higher auxiliary-electric-heater energy than the fleet average. These associations suggest that context-adjusted performance evaluation can help identify installations warranting further investigation using routinely available operational data, without requiring prior fault diagnosis.

As heat-pump deployment increases, such context-adjusted evaluation could support scalable, low-cost performance screening, help installers prioritize inspections, and provide homeowners with more meaningful efficiency assessments. It could also support commissioning quality assurance without requiring detailed manual assessment of every installation. The method is intended for identifying potential performance shortfalls, not diagnosing faults or determining optimal operating conditions. Further validation across manufacturers and independently verified fault cases is needed before diagnostic or regulatory use.

5 Methods

5.1 Data

To develop and evaluate the method for expected efficiency, we use three complementary field-reference datasets from Switzerland, the United Kingdom, and Germany, comprising 176 heat pumps. The German reference dataset comprises 67 quality-screened systems selected from a substantially larger manufacturer telemetry fleet. The complete manufacturer fleet of 10,087 heat pumps, including these 67 reference systems, is subsequently used for temporal robustness assessment and fleet-scale application without further recalibration. The remaining fleet systems do not contribute to model development or parameter estimation. Table 2 summarizes the four datasets and their roles; extended selection and measurement details are provided in Appendix A1.1.

Table 2: Overview of datasets used for reference development, validation, and fleet-scale application.

Dataset	Systems	System-months	Quality gate	Analytical role
Swiss field-monitoring study	5	180	Participation in an FH OST field study with on-site visits to selected installations, suggesting greater attention to system installation and operation.	Reference development and validation
Open-source field monitoring	104	1,029	Publicly documented installation and performance data, with named installers, potentially encouraging greater attention to installation quality.	Reference development and validation
Quality-screened manufacturer subset	67	464	Installations by teams ranking in the best 20% for quality-control rejection rate, rework rate, and reported cold starts.	Reference development and validation
Total reference dataset	**176**	**1,673**	–	–
Large-scale manufacturer fleet	10,087	112,649	No installer-quality restriction.	fleet-scale application

5.1.1 Reference datasets

Field-monitoring in Switzerland The first dataset originates from an official field-measurement study conducted in Switzerland [44]. The installations were monitored within a dedicated research campaign involving on-site assessments and dedicated measurement equipment. Five heat pumps meet the inclusion criteria of being variable-capacity air-to-water systems with corresponding COP values available in the KEYMARK database [15]. Data is available between September 2019 to August 2022. The measurement equipment was calibrated at the Heat Pump Test Center (WPZ) in Buchs and subsequently verified on-site [45].

Open-source field-monitoring data The second dataset was obtained from `HeatpumpMonitor.org`, an OpenEnergyMonitor community platform for sharing and comparing real-world heat pump performance data. The systems are voluntarily equipped with comparatively detailed monitoring and are publicly documented by their owners and installers. HeatpumpMonitor.org uses independent heat and electricity meters, including MID-certified equipment [7, 46]. Systems were excluded when external control strategies were reported, electricity-metering boundaries were incompatible, or backup electric heating could not be measured separately. Only months containing measurements for at least 28 distinct days were retained. After filtering, 104 systems with 1,029 valid monthly observations remained.

Manufacturer field-monitoring data The manufacturer reference subset was defined using installation-team quality indicators that were independent of the heat-pump efficiency outcomes analysed in this study. Teams were retained only if they simultaneously ranked within the best 20% (that is, the lowest values) for three internally monitored quality metrics: Installation Quality Control Proof rejection rate, rework rate, and the number of reports classified internally as “cold start” per installed system. Three installation teams met the criterion; after applying the telemetry-completeness and model-input

requirements, 67 systems remained. Rather than being independently metered, thermal and electrical energy values are estimated by the heat-pump controllers using operational measurements [17]. The performance metrics encompass the electricity consumption of the compressor, fan, and auxiliary electric heater, approximately corresponding to the H3 system boundary [8, 47].

The three datasets are complementary. The Swiss study provides data from a closely supervised field-monitoring campaign, the open-source dataset offers a comparatively large and heterogeneous sample dominated by systems operating at relatively low supply temperatures, and the manufacturer dataset represents the type of telemetry that can be collected at scale without additional measurement hardware, primarily from retrofit buildings. Differences in sensor quality, measurement boundaries, temporal resolution, and data processing across the datasets must be considered when interpreting the results.

The three reference datasets were selected to represent the performance that residential air-to-water heat pumps (AWHPs) can reasonably achieve under real-world operating conditions. While optimal installation and operation cannot be guaranteed, each dataset provides grounds for confidence in installation quality. The Swiss systems participated in a structured FH OST field study involving on-site visits to selected installations, suggesting greater attention to installation and operation. The HeatpumpMonitor.org systems publicly disclose installation details and operational performance, often alongside installer identities, creating reputational incentives for installation quality. The manufacturer subset comprises installations by three teams that simultaneously ranked within the best 20% for quality-control rejection rate, rework rate, and reported cold starts. Together, these selection criteria reduce the likelihood that systematic installation deficiencies influence the empirical field reference, without assuming that the selected systems operate optimally.

5.1.2 Large-scale manufacturer fleet

The final method was applied to the manufacturer telemetry dataset comprising 10,087 residential air-to-water heat pumps monitored from 1 April 2025 to 1 April 2026. The systems originate from a single manufacturer and are installed predominantly in Germany, with a smaller proportion located in other European countries. Measurements are obtained from sensors integrated into the heat pumps and transmitted through their internet-connected control systems. The energy balance follows the H3 monitoring boundary of the SEPEMO taxonomy [47]; electrical and thermal energy quantities represent controller-derived estimates rather than measurements from independent energy meters. The present study uses daily features derived from the high-frequency measurements; days containing incomplete measurements or relevant data gaps are excluded.

Service records Service records from the same heat pump fleet were available for the period from 1 November 2025 to 28 February 2026. The service records were linked to the monitoring data using an anonymized system identifier. Repeated records belonging to the same heat pump were aggregated to derive a binary indicator denoting whether at least one service contact had been recorded, while the total number of recorded contacts per system was retained as a complementary count outcome. In the final common analysis cohort, 9,604 systems had the required monitoring variables, of which 1,017 systems (10.6%) had at least one recorded service contact. The records do not distinguish between technical faults and other service-related interactions and are interpreted as an indicator of recorded service demand rather than an exhaustive fault history.

5.2 Estimation of expected performance under observed operating conditions

The aim of the study is to estimate the seasonal performance factor that can reasonably be expected from an individual heat pump under a specified set of operational conditions. These conditions include the relevant outdoor and supply temperatures, heat pump model characteristics, and contributions of CH and DHW. Figure 1 illustrates the idea of progression from thermodynamic potential to achieved field performance. The Carnot COP represents the physical efficiency limit imposed by the temperature lift. This potential is first adjusted by the laboratory factor F_{lab}, derived from model-specific COP data, and subsequently by the empirical field factor F_{field}, which accounts for the systematic transition from laboratory to field performance. The resulting context-specific expected SPF is subsequently compared with the achieved SPF to quantify target achievement. Performance is evaluated at monthly resolution, the finest common resolution across the reference datasets, to balance short-term operating variability with timely assessment. While daily efficiency is sensitive to changing weather and operating conditions,

annual evaluation delays feedback. Previous research has demonstrated the feasibility of multi-day seasonal performance estimation [12], while monthly performance indicators are well established in field-monitoring studies [10, 14]. For central heating, observations from May to September were excluded, as were months with a mean outdoor temperature of at least 20 °C or less than 50 kWh of CH electrical input from the compressor and auxiliary heater. DHW observations were retained throughout the year.

5.2.1 Laboratory-based performance expectation

Thermodynamic reference. For each observation period i and operating mode $m \in \{\mathrm{CH}, \mathrm{DHW}\}$, the effect of the temperature lift is represented by the Carnot coefficient of performance,

$$COP_{\mathrm{Carnot},i,m} = \frac{T_{\mathrm{sup},i,m} + 273.15}{(T_{\mathrm{sup},i,m} + 273.15) - (T_{\mathrm{out},i} + 273.15)}, \tag{2}$$

where $T_{\mathrm{out},i}$ denotes the mean outdoor temperature recorded by the outdoor sensor during observation period i, and $T_{\mathrm{sup},i,m}$ denotes the corresponding mean supply temperature measured during CH or DHW operation. These directly measured temperatures were used without applying additional offsets to approximate refrigerant evaporation or condensation temperatures. Such adjustments depend on heat exchanger characteristics and operating conditions, introducing assumptions that may not be consistently applicable across different heat pump models and monitoring systems [6, 16]. Using measured outdoor and supply temperatures provides a transparent and reproducible thermodynamic reference based on commonly available signals. Temperatures are converted from °C to kelvin for the Carnot calculation. Period-mean temperatures were used as the finest common temporal resolution across the three reference datasets, allowing expected and achieved performance to be evaluated over identical periods. Owing to the nonlinearity of the Carnot relationship, this represents a period-level approximation rather than a reconstruction of instantaneous thermodynamic performance.

Model-specific laboratory performance. Laboratory COP values were available for the outdoor-temperature test points at -7°C, 2°C, and 7°C each at supply temperatures of 35°C and 55°C. To estimate laboratory performance at the supply temperature observed in the field, the laboratory COP values were first expressed relative to the corresponding Carnot COP.

For each laboratory outdoor-temperature point $T_j \in \{-7, 2, 7\}$ °C, the Carnot-normalized laboratory efficiencies at supply temperatures of 35 and 55°C are

$$\eta_{\mathrm{lab},j,35} = \frac{COP_{\mathrm{lab}}(T_j, 35)}{COP_{\mathrm{Carnot}}(T_j, 35)}, \tag{3}$$

and

$$\eta_{\mathrm{lab},j,55} = \frac{COP_{\mathrm{lab}}(T_j, 55)}{COP_{\mathrm{Carnot}}(T_j, 55)}. \tag{4}$$

The Carnot-normalized laboratory efficiency corresponding to the observed supply temperature was then obtained by linear interpolation or limited extrapolation:

$$\eta_{\mathrm{lab},j}\left(T^*_{\mathrm{sup}}\right) = \eta_{\mathrm{lab},j,35} + \frac{T^*_{\mathrm{sup}} - 35}{55 - 35}\left(\eta_{\mathrm{lab},j,55} - \eta_{\mathrm{lab},j,35}\right), \tag{5}$$

where

$$T^*_{\mathrm{sup}} = \min\left(70, \max\left(30, T_{\mathrm{sup}}\right)\right). \tag{6}$$

Thus, laboratory performance is interpolated within the measured range of 35-55°C, while limited linear extrapolation is allowed between 30 and 70°C. Supply temperatures outside this range are evaluated at the respective boundary to avoid extrapolation into operating conditions insufficiently represented by the available laboratory data.

Weighting of laboratory test points. Laboratory COP values were consistently available at outdoor temperatures of -7 °C, 2 °C, and 7 °C. To derive a laboratory reference representative of each observation period, we approximated its outdoor-temperature distribution by a normal distribution centered on the observed mean outdoor temperature $\mu_{T_{\mathrm{out}},i}$, as shown in Fig. 6a:

$$p_{T,i}(T) = \mathcal{N}\left(T; \mu_{T_{\mathrm{out}},i}, \sigma_T^2\right). \tag{7}$$

A fixed standard deviation of 3 K was used for both CH and DHW as a common smoothing assumption across datasets and observation periods, rather than a location-specific estimate of climatic temperature variability.

For CH, the temperature distribution was additionally weighted by relative space-heating demand, approximated according to the heating-degree principle as

$$Q_{\mathrm{rel}}(T) = \max\left(0, T_{\mathrm{bal}} - T\right), \tag{8}$$

where the heating-balance temperature T_{bal} was set to 16 °C, as shown in Fig. 6b. The resulting demand-weighted distribution assigns greater importance to colder outdoor temperatures when weighting the three laboratory test points (Fig. 6c).

Following the heating-degree principle, heating demand was approximated as proportional to the positive difference between the heating-balance temperature and outdoor temperature,

$$Q_{\mathrm{rel}}(T) = \max\left(0, T_{\mathrm{bal}} - T\right), \tag{9}$$

and the resulting heating-energy-weighted temperature distribution is

$$p_{\mathrm{CH},i}(T) = \frac{p_{T,i}(T) Q_{\mathrm{rel}}(T)}{\displaystyle\int_{-\infty}^{\infty} p_{T,i}(T) Q_{\mathrm{rel}}(T)\,\mathrm{d}T}. \tag{10}$$

For DHW, no space-heating-demand weighting was applied because DHW production is not governed by the building's space-heating demand. Consequently,

$$p_{\mathrm{DHW},i}(T) = p_{T,i}(T). \tag{11}$$

Each outdoor temperature was mapped to the nearest available laboratory test temperature. With laboratory COP values available at -7°C, 2°C, and 7°C, the corresponding assignment boundaries are given by the midpoints -2.5°C and 4.5°C. This defines the temperature regions

$$\mathcal{R}_{-7} = (-\infty, -2.5), \qquad \mathcal{R}_2 = [-2.5, 4.5), \qquad \mathcal{R}_7 = [4.5, \infty). \tag{12}$$

For operating mode m, the weight assigned to laboratory point j is

$$w_{i,m,j} = \int_{\mathcal{R}_j} p_{m,i}(T)\,\mathrm{d}T, \qquad j \in \{-7, 2, 7\}, \tag{13}$$

with

$$w_{i,m,-7} + w_{i,m,2} + w_{i,m,7} = 1. \tag{14}$$

The model- and context-specific laboratory factor is then calculated as the weighted mean fraction of Carnot performance across the three laboratory operating points:

$$F_{\mathrm{lab},i,m} = \sum_{j \in \{-7,2,7\}} w_{i,m,j}\, \eta_{\mathrm{lab},j}\left(T^*_{\mathrm{sup},i,m}\right). \tag{15}$$

The product $COP_{\mathrm{Carnot},i,m} F_{\mathrm{lab},i,m}$ therefore represents the performance expected from the specific heat pump model based on its laboratory characteristics and the temperature conditions of observation period i and operating mode m. As the laboratory weights are derived from the observed mean outdoor temperature of each period, the formulation does not rely on a prescribed climatic profile, such as the EN 14825 average climate, or location-specific design temperatures. It can therefore be applied across different climatic conditions, although transferability of the empirical field correction requires separate validation. This laboratory-based expectation does not yet account for systematic differences between standardized laboratory and field operation.

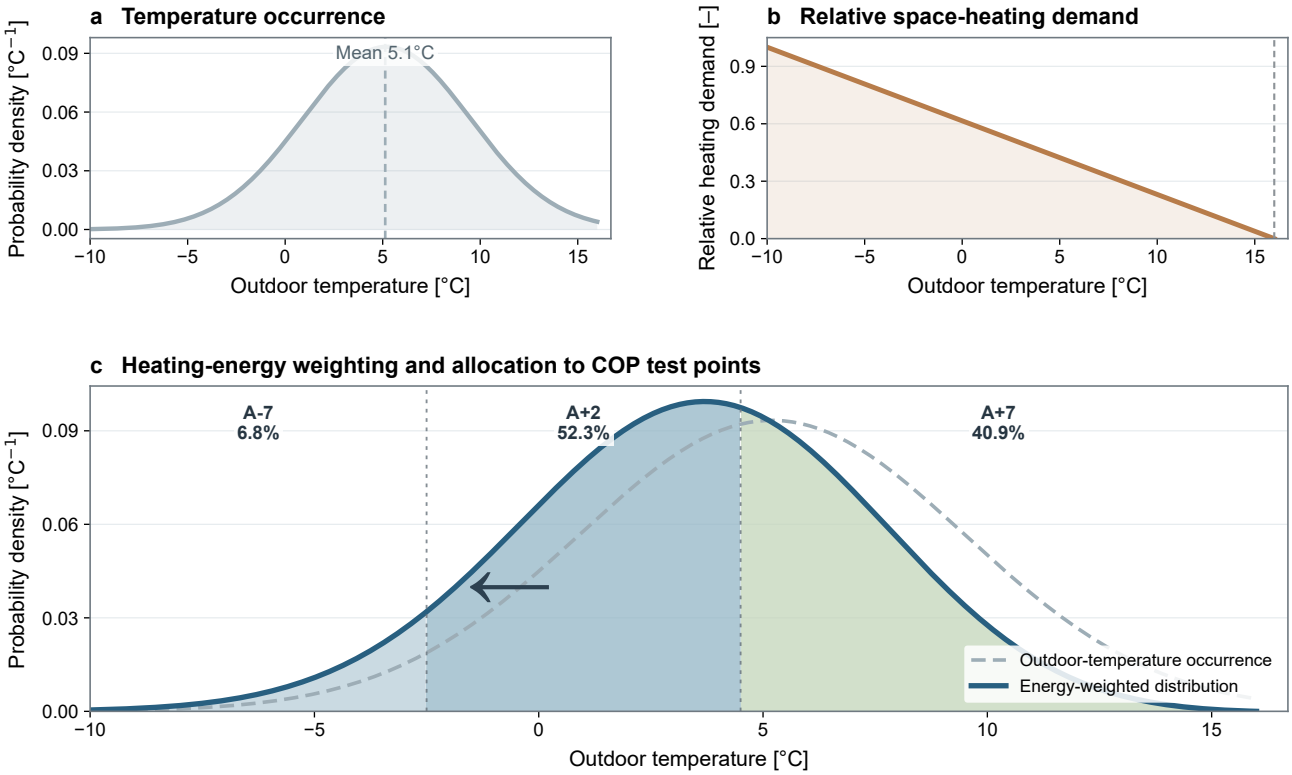


Figure 6: Illustration of the observation-period-specific weighting of laboratory COP test points for central heating. The outdoor-temperature distribution is centered on the observed mean outdoor temperature and weighted according to relative central-heating demand. Each outdoor temperature is assigned to the nearest available laboratory test outdoor temperature of -7°C, 2°C, or 7°C. The illustrated example assumes a mean outdoor temperature of 5.1°C and a standard deviation of 3.0 K.

5.2.2 Empirical field correction and model selection

Laboratory-based performance does not capture all differences between standardized testing and real-world operation. These differences may reflect cycling, part-load operation, auxiliary-electric consumption, control settings, hydraulic integration, or other installation- and operation-specific effects. An empirical field correction was therefore estimated from the curated field-reference datasets.

For each observation i and operating mode m, the observed multiplicative field factor is defined as

$$F_{\mathrm{field,obs},i,m} = \frac{SPF_{\mathrm{achieved},i,m}}{COP_{\mathrm{Carnot},i,m} F_{\mathrm{lab},i,m}}. \tag{16}$$

The field factor quantifies the remaining difference between achieved field performance and the corresponding laboratory-based expectation after accounting for temperature lift and model-specific laboratory performance. It should not be interpreted as the average field loss of the complete installed heat pump population or as a thermodynamic optimum, rather, it represents the empirical relationship observed in the curated field-reference datasets used for model development.

The nine candidate formulations included global additive and multiplicative corrections and corrections based on outdoor temperature, supply temperature, supply-temperature variability, and system utilization (Appendix Table A1). The Carnot and laboratory components were identical across all formulations, and all models were evaluated separately for CH and DHW.

Candidate models were evaluated using 20 repetitions of five-fold cross-validation grouped by heat pump system. All observations belonging to the same heat pump were assigned exclusively to either the training or validation set within a given fold, preventing information leakage between repeated observations of the same installation.

Within each repetition, predictions from the five held-out folds were pooled for evaluation. System-level fold stratification and performance aggregation are detailed in Appendix A1.3.

All quantities required by a candidate model, including field-correction parameters, temperature-bin definitions, and fallback values, were estimated exclusively from the training systems of the respective fold and applied unchanged to the held-out systems.

Predictive performance was assessed primarily using mean error (ME), mean absolute error (MAE), MdAPE, root mean square error (RMSE), coefficient of determination (R^2), and calibration intercept and slope were used as complementary measures of predictive accuracy and calibration. Mean and dispersion of target achievement were additionally examined to identify systematic over- or underprediction. Model selection considered both predictive performance and model parsimony.

Based on this model-selection procedure, an outdoor-temperature-dependent multiplicative field correction was selected for both CH and DHW. For this formulation, observations are grouped into 5°C outdoor-temperature intervals. Within each interval b, the field factor is estimated from the training observations as

$$F_{\text{field},m,b} = \frac{1}{n_{m,b}} \sum_{i \in \mathcal{B}_{m,b}} F_{\text{field,obs},i,m}, \tag{17}$$

where $\mathcal{B}_{m,b}$ denotes the training observations of operating mode m assigned to outdoor-temperature interval b, and $n_{m,b}$ denotes the corresponding number of observations.

Expected performance for observation i is then

$$SPF_{\text{expected},i,m} = COP_{\text{Carnot,i,m}} \cdot F_{\text{lab,i,m}} \cdot F_{\text{field},b,m}. \tag{18}$$

The outermost temperature bins were extended beyond the training-data range; fallback and numerical safeguards are documented in Appendix A1.3.

After model selection, the selected formulation was fitted once to the complete curated field-reference datasets separately for CH and DHW. The resulting parameters were fixed before application to the large-scale manufacturer fleet. Apart from the predefined 67-system quality-screened subset included in the reference sample, no additional systems from the large-scale manufacturer fleet contributed to model development or parameter estimation.

5.2.3 Expected total performance and target achievement

The achieved SPF of operating mode m over the considered observation period is calculated as

$$SPF_{\text{achieved},m} = \frac{Q_{\text{out},m}}{E_{\text{in},m}}, \tag{19}$$

where $Q_{\text{out},m}$ denotes useful heat output and $E_{\text{in},m}$ denotes the corresponding electrical energy input over the same observation period.

Total expected performance was not fitted as a separate model. Instead, the independently estimated CH and DHW expectations were combined according to their contributions to useful heat output:

$$SPF_{\text{expected,total}} = \left(\frac{1-y}{SPF_{\text{expected,CH}}} + \frac{y}{SPF_{\text{expected,DHW}}} \right)^{-1}, \tag{20}$$

where

$$y = \frac{Q_{\text{out,DHW}}}{Q_{\text{out,DHW}} + Q_{\text{out,CH}}} \tag{21}$$

denotes the fraction of total useful heat output attributable to DHW production. Consequently, $1-y$ denotes the corresponding CH fraction.

The achieved total SPF is calculated directly from the combined useful heat output and electrical energy input:

$$SPF_{\text{achieved,total}} = \frac{Q_{\text{out,CH}} + Q_{\text{out,DHW}}}{E_{\text{in,CH}} + E_{\text{in,DHW}}}. \tag{22}$$

For CH, DHW, and total performance, target achievement is defined as

$$TA_m = \frac{SPF_{\text{achieved},m}}{SPF_{\text{expected},m}} \cdot 100\%. \tag{23}$$

A target achievement of 100% indicates agreement between achieved performance and the context-specific expectation derived from the curated field-reference datasets. Values below 100% indicate performance below this reference, whereas values above 100% indicate performance exceeding it. Target achievement does not represent thermodynamic optimality and cannot identify suboptimal system parameterization, such as unnecessarily high heating-curve settings, or specific installation, control, or component faults. Rather, it serves as a screening metric for identifying systems whose achieved performance deviates substantially from the level expected under their observed operating conditions.

5.3 Evaluation and fleet-scale application

After model selection and final parameter estimation, the approach was evaluated for out-of-sample robustness across systems, monitoring sources, and observation periods, and subsequently applied without further calibration to the large-scale manufacturer fleet.

5.3.1 Out-of-sample robustness and transportability

Predictive performance and calibration. Prediction accuracy was evaluated using the system-grouped out-of-fold predictions described above. Metrics used for model selection were calculated for each complete out-of-fold prediction vector and summarized across 20 repetitions. For secondary analyses, the held-out predictions for each observation were averaged across repetitions; every contributing prediction came from a model fitted without that system's data. Calibration was assessed using the intercept and slope of a linear regression of achieved against expected SPF. Variation in target achievement (TA) was summarized using the standard deviation across observations, the standard deviation of system-specific mean TA, and the pooled within-system standard deviation of observations around their respective system means. These measures describe variability rather than formal variance components. Residual dependence on outdoor temperature, supply temperature, DHW share, auxiliary-electric-heater share, and specific energy output was assessed using separate univariate linear regressions for CH, DHW, and total performance. The calculations and diagnostics are detailed in Appendix A2.7.

Cross-dataset transportability. Transportability across the three reference datasets was assessed using leave-one-dataset-out evaluation, separately for CH and DHW. In each iteration, one complete data source was withheld, the parameters of the selected model formulation were re-estimated using the remaining two sources, and the fitted parameters were applied without recalibration to the withheld source. The procedure was repeated until each dataset had served once as the held-out source; no system identifiers overlapped between training and held-out datasets. Prediction accuracy, calibration, and the distribution of TA were evaluated for each withheld dataset. The results were not used to modify the model form, final parameters, or screening thresholds. Because all three sources contributed to the preceding candidate-model comparison, this analysis evaluates transportability of the selected model structure rather than fully independent external validation of model-form selection.

5.3.2 Robustness across observation periods

To assess sensitivity to temporal aggregation, the field-correction parameters estimated from monthly reference observations were applied without re-estimation to yearly aggregated observations from the reference datasets and to both monthly and yearly observations from the manufacturer fleet. At each resolution, expected and achieved SPF were calculated using the corresponding operating temperatures, mode-specific useful heat contributions, and accumulated energy quantities. CH and DHW expectations were calculated separately and combined according to Eq. 20. Prediction accuracy and TA distributions were compared across observation periods; the reported metrics are specified in Appendix A2.7.

5.3.3 Fleet screening and performance groups

The final model was applied without further calibration to annual system-level observations from the manufacturer fleet. Screening combined a TA threshold of 80% with an absolute achieved-SPF threshold of 3.0. The relative threshold identifies performance more than 20% below the context-specific expectation; the absolute threshold provides a policy-relevant reference corresponding to the minimum calculated design-stage Jahresarbeitszahl (EN: seasonal performance factor) (JAZ) previously specified in the German

Federal Funding for Efficient Buildings programme [18]. Design-stage JAZ and measured operational SPF are not directly equivalent, and neither threshold represents a technical fault boundary.

Threshold sensitivity was assessed by varying TA from 70% to 90% in five-percentage-point increments and achieved SPF from 2.0 to 3.5 in increments of 0.25. For each combination, the proportion of valid systems below both thresholds, and the proportion below the absolute SPF threshold alone, were calculated separately for CH, DHW, and total performance.

Performance-group definitions. Among systems with TA below 80%, Zone I comprises systems with SPF_{achieved} <3.0 and SPF_{expected} ≥3.0; Zone II comprises systems with SPF_{achieved} <3.0 and SPF_{expected} <3.0; and Zone III comprises systems with SPF_{achieved} ≥3.0 and SPF_{expected} ≥3.0. All remaining systems are classified as Rest. Zones I and II satisfy both the relative and absolute screening criteria, whereas Zone III identifies substantial underperformance despite an achieved SPF of at least 3.0.

Operating characteristics of the groups. Annual system-level operating characteristics were compared descriptively across the groups defined by total SPF. For each variable, the mean within a group was divided by the corresponding mean across the complete valid fleet; a ratio of 1.0 indicates the fleet mean. Compressor-alarm duration was calculated as the mean monthly accumulated duration of the compressor alarm state, averaged over available monitoring months for each system. These associations are not interpreted as causal evidence of specific faults.

Association with recorded service activity. Service records were linked to the monitoring data by anonymized system identifier and aggregated per system. Systems were compared using TA below versus at least 80% and achieved SPF below versus at least 3.0. Service-contact incidence was the fraction of systems with at least one recorded contact; 95% Wilson score confidence intervals were calculated for these proportions. The number of contacts per system was analysed using separate negative-binomial generalized linear models with a log link and performance group as the explanatory variable, to account for overdispersion. Group-specific mean contact frequencies and 95% confidence intervals were obtained from the fitted models. For the service-association analysis, achieved SPF and target achievement were evaluated over the month preceding the corresponding service-observation period, rather than over the full annual monitoring period used for fleet classification. This temporal alignment allows performance before recorded service activity to be examined, while avoiding the inclusion of subsequent operating periods in which performance may have changed. The resulting classifications can therefore differ from those obtained using annual performance.

References

1. International Energy Agency (IEA). *Heat Pumps* Accessed: 2023-11-28 (International Energy Agency (IEA), Paris, 2023). https://www.iea.org/energy-system/buildings/heat-pumps.
2. International Energy Agency. *The Future of Heat Pumps* (International Energy Agency, Paris, 2022). https://www.iea.org/reports/the-future-of-heat-pumps.
3. International Energy Agency. *Heat Pump Monitor 2026* 2026. https://www.iea.org/reports/heat-pump-monitor-2026.
4. Weigert, A., Hopf, K., Günther, S. A. & Staake, T. Heat pump inspections result in large energy savings when a pre-selection of households is performed: A promising use case of smart meter data. *Energy Policy* **169,** 113156. doi:10.1016/j.enpol.2022.113156 (2022).
5. Fraunhofer ISE. *Wärmepumpen in Bestandsgebäuden: Ergebnisse aus dem Forschungsprojekt WPsmart im Bestand* tech. rep. (Fraunhofer-Institut für Solare Energiesysteme ISE, Freiburg, Germany, 2020). https://www.ise.fraunhofer.de/content/dam/ise/de/downloads/pdf/Forschungsprojekte/BMWi-03ET1272A-WPsmart_im_Bestand-Schlussbericht.pdf.
6. Fraunhofer ISE. *Wärmepumpen in Bestandsgebäuden: Abschlussbericht des Projekts WP-QS im Bestand* tech. rep. Project period 2019–2024 (Fraunhofer-Institut für Solare Energiesysteme ISE, Freiburg, Germany, 2025). https://www.ise.fraunhofer.de/de/forschungsprojekte/wp-qs-im-bestand.html.
7. Rosenow, J., Lea, T. & Boni, G. Bridging the efficiency divide: open-source insights into UK heat pump performance gaps. *Energy and Buildings* **352,** 116785. doi:10.1016/j.enbuild.2025.116785 (2026).
8. European Committee for Standardization. *Air conditioners, liquid chilling packages and heat pumps, with electrically driven compressors, for space heating and cooling, commercial and process cooling - Testing and rating at part load conditions and calculation of seasonal performance* Standard EN 14825:2022 (CEN, Brussels, Belgium, 2022).
9. Verein Deutscher Ingenieure. *VDI 4650 Part 1: Calculation of the Seasonal Performance Factor of Heat Pump Systems* Düsseldorf, Germany, 2019.
10. Miara, M., Günther, D., Langner, R., Helmling, S. & Wapler, J. 10 Years of Heat Pumps Monitoring in Germany: Outcomes of Several Monitoring Campaigns. From Low-Energy Houses to Un-Retrofitted Single-Family Dwellings. In *Proceedings of the 12th IEA Heat Pump Conference* Conference paper K.1.5.1 (Rotterdam, Netherlands, 2017). https://heatpumpingtechnologies.org/publications/k-1-5-1-10-years-of-heat-pumps-monitoring-in-germany-outcomes-of-several-monitoring-campaigns-from-low-energy-houses-to-un-retrofitted-single-family-dwellings/.
11. Bernal, S., Berthold, M., Bertsch, S. & Eschmann, M. *Feldmessungen von Wärmepumpen-Anlagen: Heizsaison 2022/23* Jahresbericht im Auftrag von EnergieSchweiz, Bundesamt für Energie (OST – Ostschweizer Fachhochschule, 2023). https://pubdb.bfe.admin.ch/de/publication/download/11629.
12. Sun, X., Wang, Z., Li, X., Xu, Z., Yang, Q. & Yang, Y. Seasonal heating performance prediction of air-to-water heat pumps based on short-term dynamic monitoring. *Renewable Energy* **180,** 829–837. doi:10.1016/j.renene.2021.08.130 (2021).
13. Brudermueller, T., Potthoff, U., Fleisch, E., *et al.* Estimation of energy efficiency of heat pumps in residential buildings using real operation data. *Nature Communications* **16,** 2834. doi:10.1038/s41467-025-58014-y (2025).
14. Fraga, C., Mermoud, F., Hollmuller, P., Pampaloni, E. & Lachal, B. Large solar driven heat pump system for a multifamily building: Long term in-situ monitoring. *Solar Energy* **114,** 427–439. doi:10.1016/j.solener.2015.01.039 (2015).
15. Heat Pump KEYMARK. *Certified Heat Pump Products* https://keymark.eu/en/products/heatpumps/certified-products. Accessed: 23 July 2026.
16. OpenEnergyMonitor. *HeatpumpMonitor.org: % of Ideal Carnot COP* Accessed 26 October 2026. 2026. https://docs.openenergymonitor.org/heatpumpmonitor/prc_carnot.html.
17. Nolting, L., Steiger, S. & Praktiknjo, A. Assessing the validity of European labels for energy efficiency of heat pumps. *Journal of Building Engineering* **18,** 476–486. doi:10.1016/j.jobe.2018.02.013 (2018).

18. Bundesministerium für Wirtschaft und Klimaschutz. *Richtlinie für die Bundesförderung für effiziente Gebäude – Einzelmaßnahmen (BEG EM)* tech. rep. BAnz AT 30.12.2022 B1. Vom 9. Dezember 2022. Veröffentlicht im Bundesanzeiger am 30. Dezember 2022 (Bundesministerium für Wirtschaft und Klimaschutz, 2022). https://www.bundesanzeiger.de/pub/publication/xSizk6DUlWm93L4XrkY/content/xSizk6DUlWm93L4XrkY/BAnz%20AT%2030.12.2022%20B1.pdf?inline.

19. O'Hegarty, R., Kinnane, O., Lennon, D. & Colclough, S. Air-to-water heat pumps: Review and analysis of the performance gap between in-use and product rated performance. *Renewable and Sustainable Energy Reviews* **155,** 111887. doi:10.1016/j.rser.2021.111887 (2022).

20. Carroll, P., Chesser, M. & Lyons, P. Air Source Heat Pumps field studies: A systematic literature review. *Renewable and sustainable energy reviews* **134,** 110275. doi:10.1016/j.rser.2020.110275 (2020).

21. Bernal, S., Brendel, L., Arpagaus, C. & Bertsch, S. Steady-State-Detektion zur Beurteilung von Wärmepumpen: Eine Analyse von Feldmessdaten. In *Deutsche Kälte- und Klimatagung 2024* Conference paper AA IV 07 (Dresden, Germany, 2024). https://www.researchgate.net/publication/389055506_Steady-State-Detektion_zur_Beurteilung_von_Warmepumpen_Eine_Analyse_von_Feldmessdaten.

22. Brudermueller, T., Kreft, M., Fleisch, E. & Staake, T. Large-scale monitoring of residential heat pump cycling using smart meter data. *Applied Energy* **350,** 121734. doi:10.1016/j.apenergy.2023.121734 (2023).

23. Lämmle, M., Metz, J., Kropp, M., Wapler, J., Oltersdorf, T., Günther, D., Herkel, S. & Bongs, C. Heat Pump Systems in Existing Multifamily Buildings: A Meta-Analysis of Field Measurement Data Focusing on the Relationship of Temperature and Performance of Heat Pump Systems. *Energy Technology,* 2300379. doi:10.1002/ente.202300379 (2023).

24. Tejeda, A., Milu, A., Riviere, P. & Marchio, D. Energy Consequences of Non-optimal Heat Pump Parameterization. In *International High Performance Buildings Conference* (2014). https://docs.lib.purdue.edu/ihpbc/110/.

25. Bagarella, G., Lazzarin, R. & Noro, M. Sizing strategy of on-off and modulating heat pump systems based on annual energy analysis. *International Journal of Refrigeration* **65,** 183–193. doi:10.1016/j.ijrefrig.2016.02.015 (2016).

26. Dongellini, M., Naldi, C. & Morini, G. L. Sizing effects on the energy performance of reversible air-source heat pumps for office buildings. *Applied Thermal Engineering* **114,** 1073–1081. doi:10.1016/j.applthermaleng.2016.12.010 (2017).

27. Caird, S., Roy, R. & Potter, S. Domestic heat pumps in the UK: user behaviour, satisfaction and performance. *Energy Efficiency* **5,** 283–301. doi:10.1007/s12053-012-9146-x (2012).

28. De Wilde, M. A Heat Pump Needs a Bit of Care: On Maintainability and Repairing Gender–Technology Relations. *Science, Technology, & Human Values* **46,** 1261–1285. doi:10.1177/0162243920978301 (2021).

29. Oikonomou, E., Zimmermann, N., Davies, M. & Oreszczyn, T. Behavioural Change as a Domestic Heat Pump Performance Driver: Insights on the Influence of Feedback Systems from Multiple Case Studies in the UK. *Sustainability* **14,** 16799. doi:10.3390/su142416799 (2022).

30. International Energy Agency - Technology Collaboration Programme on Heat Pumping Technology (IEA - HPT TCP). *Annex 56 on Digitalization and IoT for Heat Pumps* Accessed: 2023-12-18 (International Energy Agency (IEA), Sweden, 2023). https://heatpumpingtechnologies.org/annex56/.

31. Song, Y., Rolando, D., Marchante Avellaneda, J., Zucker, G. & Madani, H. Data-driven soft sensors targeting heat pump systems. *Energy Conversion and Management* **279,** 116769. doi:10.1016/j.enconman.2023.116769 (2023).

32. Gleeson, C. P. Residential heat pump installations: the role of vocational education and training. *Building Research & Information* **44,** 394–406. doi:10.1080/09613218.2015.1082701 (2016).

33. Decuypere, R., Robaeyst, B., Hudders, L., Baccarne, B. & Van de Sompel, D. Transitioning to energy efficient housing: Drivers and barriers of intermediaries in heat pump technology. *Energy Policy* **161,** 112709. doi:10.1016/j.enpol.2021.112709 (2022).

34. Gleeson, C., Biddulph, P., Lowe, R., Love, J., Summerfield, A., Oikonomou, E., Wingfield, J. & Martin, C. Analysis of heat pump data from the renewable heat premium payment scheme (RHPP) to the Department of Business, Energy and Industrial Strategy: compliance with MCS Installation Standards. *Issued: February 2017* (2017).

35. OST – Ostschweizer Fachhochschule. *OptiPower – Untersuchung der optimalen Auslegung der Leistung von Heiz- und Kühlsystemen für Wohn- und Verwaltungsgebäude* Research project documentation. 2023. https://www.ost.ch/de/projekt/optipower-untersuchung-der-optimalen-auslegung-der-leistung-von-heiz-und-kuehlsystemen-fuer-wohn-und-verwaltungsgebaeude-new649aa37c8269a267115120.

36. Stadelmann, M. & Schubert, R. How Do Different Designs of Energy Labels Influence Purchases of Household Appliances? A Field Study in Switzerland. *Ecological Economics* **144,** 112–123. doi:10.1016/j.ecolecon.2017.07.031 (2018).

37. Ntziachristos, L., Mellios, G., Tsokolis, D., Keller, M., Hausberger, S., Ligterink, N. & Dilara, P. In-use vs. type-approval fuel consumption of current passenger cars in Europe. *Energy Policy* **67,** 403–411. doi:10.1016/j.enpol.2013.12.013 (2014).

38. Kokoni, S. & Leach, M. Policy mechanisms to support heat pump deployment: A UK case study based on techno-economic modelling. *Renewable and Sustainable Energy Transition* **1,** 100009. doi:10.1016/j.rset.2021.100009 (2021).

39. Rosenow, J., Gibb, D., Nowak, T. & Lowes, R. Heating up the global heat pump market. *Nature Energy* **7,** 901–904. doi:10.1038/s41560-022-01104-8 (2022).

40. Rosenow, J., Galvin, R., O'Mara, S. & Lowes, R. Total cost of ownership of heat pumps and policy choice: The case of Great Britain. *Iscience* **28,** 111784. doi:10.1016/j.isci.2025.111784 (2025).

41. Ryland, M. & He, W. Heating economics evaluated against emissions: An analysis of low-carbon heating systems with spatiotemporal and dwelling variations. *Energy and Buildings* **277,** 112561. doi:10.1016/j.enbuild.2022.112561 (2022).

42. Brudermueller, T., Wirth, F., Weigert, A. & Staake, T. Automatic Differentiation of Variable and Fixed Speed Heat Pumps With Smart Meter Data. In *2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)* 412–418 (2022). doi:10.1109/SmartGridComm52983.2022.9961055.

43. Brudermueller, T., Breer, F. & Staake, T. Disaggregation of Heat Pump Load Profiles From Low-Resolution Smart Meter Data. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* 228–231 (Association for Computing Machinery, Istanbul, Turkey, 2023). doi:10.1145/3600100.3623731.

44. Bernal, S., Berthold, M., Bertsch, S. & Eschmann, M. *Bericht "Feldmessungen von Wärmepumpen-Anlagen Heizsaison 2022/23"* Jahresbericht. Erstellt im Auftrag von EnergieSchweiz, Bundesamt für Energie BFE (Institut für Energiesysteme, OST – Ostschweizer Fachhochschule, Rapperswil, Switzerland, 2023). https://www.ost.ch/fileadmin/dateiliste/3_forschung_dienstleistung/institute/ies/wpz/sonstige_wichtige_dokumente/2023_jahresbericht_feldmessungen.pdf.

45. Bernal, S. N., Potthoff, U. & Brendel, L. P. M. *Monitoring von Wärmepumpen: Vergleich zwischen eingebauten Komponenten und Feldmessdaten* Schlussbericht (EnergieSchweiz, Bundesamt für Energie (BFE), Bern, Switzerland, 2026).

46. OpenEnergyMonitor. *HeatpumpMonitor.org: About, Accuracy and Boundaries* Accessed 26 October 2026. 2026. https://heatpumpmonitor.org/about.

47. Zottl, A., Nordman, R., Coevoet, M., Rivière, P., Miara, M., Benou, A. & Riederer, P. *Concept for Evaluation of SPF, Version 2.2: A Defined Methodology for Calculation of the Seasonal Performance Factor and a Definition Which Devices of the System Have to Be Included in This Calculation: Heat Pumps with Hydronic Heating Systems* tech. rep. European Commission project IEE/08/776/SI2.529222 (SEPEMO-Build, 2012). https://minesparis-psl.hal.science/hal-00769989.

48. Swiderek, S., Diefenbach, N., Behem, G. & Großklos, M. *Wärmepumpen-Praxis im hessischen Wohngebäudebestand: Endbericht des Forschungsvorhabens (EN: Heat Pump Performance in Existing Residential Buildings in Hessen: Final Research Report)* tech. rep. (Institut Wohnen und Umwelt (IWU), Darmstadt, Germany, 2025). https://www.iwu.de/fileadmin/publikationen/energie/2025_IWU_Swiderek-et-al_WP-Hessen_20251014.pdf.