FINAL REPORT

AHRI Report No. 8026

Repeatability and Reproducibility Assessment of CSA EXP07:19 and AHRI 210-240:2023

Final Report
Published December 2022

Prepared By:

Parveen Dhillon, Dr. Travis W. Horton, Dr. James E. Braun

Purdue University
School of Mechanical Engineering
585 Purdue Mall
West Lafayette, IN 47907-2088

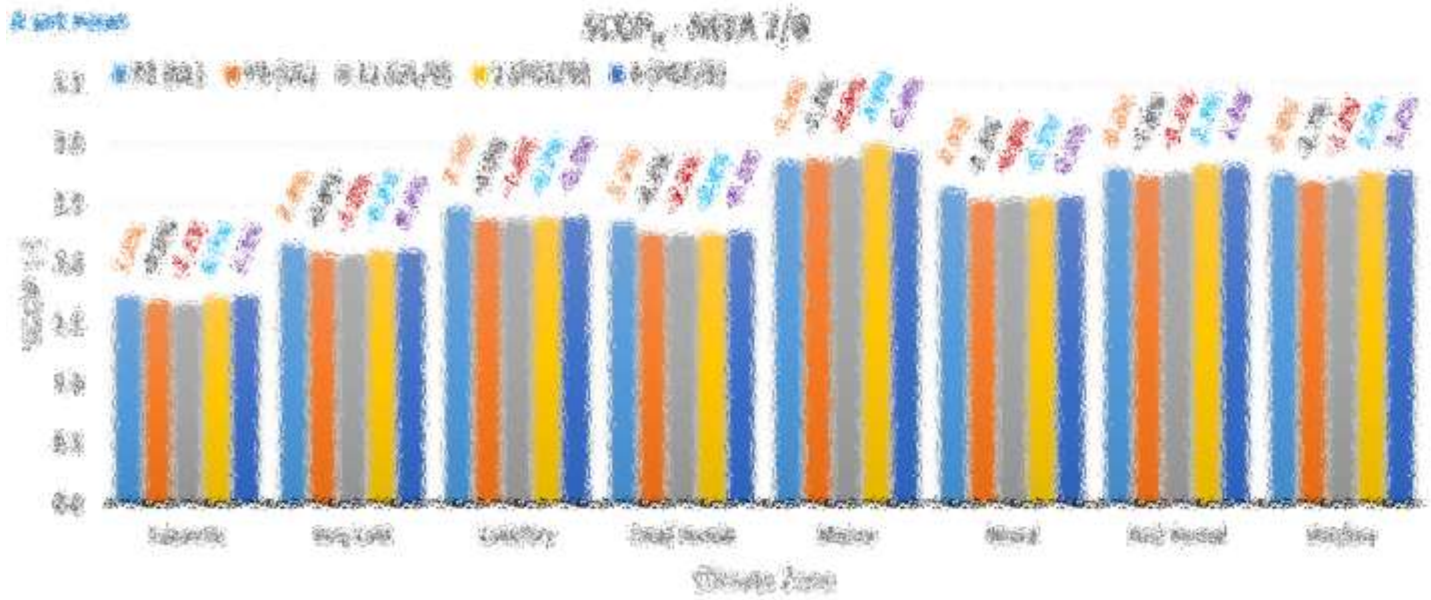
Prepared For:

Air-Conditioning, Heating and Refrigeration Institute
2311 Wilson Boulevard, Suite 400, Arlington, Virginia 22201-3001

DISCLAIMER

This report was prepared as an account of work sponsored by the Air-Conditioning, Heating, and Refrigeration Institute (AHRI). Neither AHRI, its research program financial supporters, or any agency thereof, nor any of their employees, contractors, subcontractors or employees thereof - makes any warranty, expressed or implied; assumes any legal liability or responsibility for the accuracy, completeness, any third party's use of, or the results of such use of any information, apparatus, product, or process disclosed in this report; or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute nor imply its endorsement, recommendation, or favoring by AHRI, its sponsors, or any agency thereof or their contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of AHRI, its program sponsors, or any agency thereof.

PURDUE
UNIVERSITY



AHRI 8026 - Repeatability and Reproducibility Assessment of CSA EXP07:19 and AHRI 210-240:2023

Project: AHRI 8026

Parveen Dhillon, Dr. Travis W. Horton, Dr. James E. Braun

May 24, 2022

CHPB CENTER FOR HIGH PERFORMANCE BUILDINGS AT PURDUE

RAY W. HERRICK LABORATORIES

Executive Summary

A new load-based testing methodology that forms the basis of the CSA EXP07:2019 standard draft involves measuring the performance of residential heat-pump and air-conditioning systems with their embedded controllers and thermostats in psychrometric test facilities. Before adopting a new testing and rating procedure, it is important to assess its repeatability and reproducibility. Repeatability here refers to the consistency in a test unit's measured performance when the test procedure is applied multiple times in the same lab. In contrast, reproducibility refers to the consistency in a test unit's measured performance across different test facilities. In this project, round-robin tests were conducted with multiple heat pumps of varied sizes and types in two test labs, UL and PG&E, to assess EXP07 repeatability and reproducibility. In order to compare results with the current standard testing and rating approach, round-robin tests were also conducted with some of these heat pumps based on the AHRI 210/240-2023 testing procedure. Table E.1 provides a brief description of the test units and their rated performance as per AHRI 210/240.

Table E.1 Test units' description and rated performance

Test Unit	AFS	NEEA4	NEEA7 & NEEA7B	NRCan8	NRCan10
Description	5-ton split-system ducted (single-stage)	1-ton min-split ductless (variable-speed)	3-ton split-system ducted (variable-speed)	1.5-ton min-split ductless (variable-speed)	2-ton split-system ducted (variable-speed)
Q_C [Btu/h] @ 95°F	55000	10900	33000	14000	24200
EER @ 95°F	13	12.5	12.5	13	14.5
SEER	16	20	17.8	21.6	19
Q_H [Btu/h] @ 47°F	55000	13600	38000	18000	24000
HSPF (Region IV)	8.5	12	11	11.7	10.5
Q_H [Btu/h] @ 17°F	33600	8800	29000	12100	-
Air Flow Rate [SCFM]	1600	-	-	-	-

Figure E.1 shows the progression of testing as per EXP07 and AHRI 210/240 for different test units in the two labs. Each box named either CSA EXP07 and AHRI 210/240-2023 represents one complete set of cooling and heating test intervals based on the corresponding test methodology in the test lab highlighted on the left-hand side. In this report "test interval" refers to a single test at a particular test condition and "test" refers to the entire set of test intervals as per the test methodology associated with either EXP07 or AHRI 210/240. For example, as per EXP07, one test consists of 5 cooling dry coil, 4 cooling humid coil, 6 heating continental, and 4 heating marine test intervals as shown in Table 4 and Table 5. In Figure E.1, the number in the box on the right side adjacent to each test shows the corresponding test name or number. This test name or number in combination with the test unit and test lab provide a unique identifier for each set of test intervals data. Note that while referring to test names with "PREV" in their name, a shortform "P" is generally used in the text in place of "PREV", for example, "PREV1" and "P1" refer to the same test. Figure E.1 also highlights whether a test unit was taken out of test rooms and re-installed in between different tests either in same or different lab. Further for the NEEA4 unit, test P1 data from the PG&E lab was not considered in the analysis, as at that time load-based testing was being set-up at the PG&E lab and confidence in the consistency of collected data is low. It can be seen that three units, AFS, NRCan8, and NRCan10, were only tested at UL, and the performance of NRCan8 and NRCan10 was only measured based on EXP07. As a result, the tests for these three units were only utilized to examine the repeatability of the test methodologies. Since the other two test unit models, NEEA4 and NEEA7 & NEEA7B, were tested in both labs based on both test approaches, the results were utilized to assess repeatability and reproducibility for both methods. One thing to note regarding NEEA7 is that after the initial phase of UL testing when it was shipped to PG&E, it was discovered during setup that the unit was not operating properly. As a result, a different unit but the same model, NEEA7B, was used for PG&E tests and one final test at UL. It should be noted that even though test unit was changed from NEEA7 to NEEA7B in the middle of the test sequence, the results of these two different units were still compared for overall repeatability and reproducibility assessment of EXP07 and AHRI 210/240 as the same model was utilized in testing. Further, here NEEA7(B) name is used to indicate that the analysis results being discussed include the test results of both test units NEEA7 and NEEA7B.

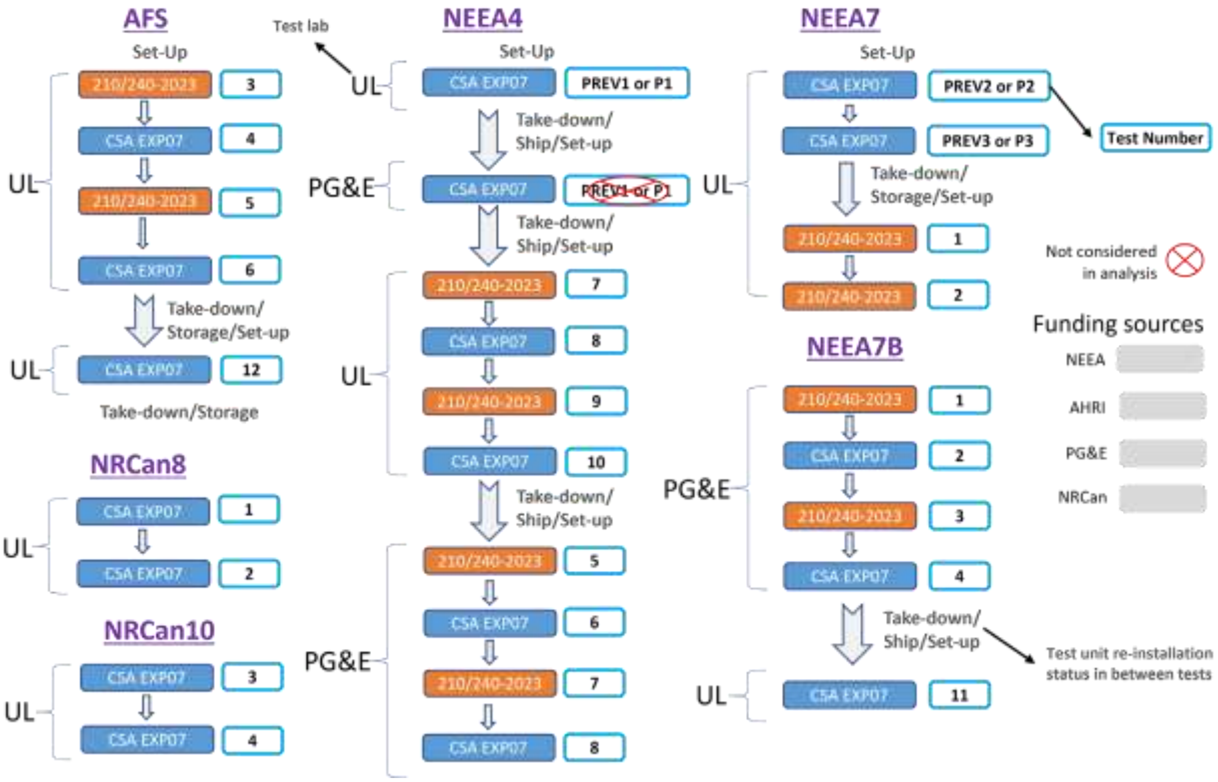


Figure E.1 Round robin tests sequence for repeatability and reproducibility analysis of CSA EXP07:19 and AHRI 210/240-2023

For CSA EXP07 and AHRI 210/240 repeatability assessment, each test unit's performance across different tests from the same lab is compared, whereas, for reproducibility assessment, NEEA4 and NEEA7(B) (NEEA7 & NEEA7B) performance from different tests across two labs are compared. For repeatability and reproducibility evaluation, different datasets are defined based on the nomenclature approach outlined in Table E.2. Each dataset name is defined based on the test unit, test lab, test methodology, number of tests in the dataset, and whether the test unit was reinstalled or not in between different tests in the dataset, either in in the same lab or different labs.

Table E.2 Dataset nomenclature approach

Test Unit –	**Test Lab** –	**Test Methodology** –	**No. of Tests in Dataset** |	**Unit Reinstallation Status between Tests**
AFS **NRCan8** **NRCan10** **NEEA4** **NEEA7** **NEEA7B** **NEEA7(B):** NEEA7 & NEEA7B	**UL** **PGE:** PG&E **UL&PGE:** UL and PG&E	**E:** EXP07-2019 **A:** AHRI 210/240-2023	**2** **3** **4** **5** (Individual tests included in dataset are provided in Table E.3)	**R:** Test unit reinstalled at least once in between different tests in the dataset, either in the same lab or in different labs **W:** Tests were performed without test unit reinstallation in between different tests in the dataset

Table E.3 shows the defined datasets for EXP07 and AHRI 210/240 repeatability and reproducibility evaluation based on the nomenclature outlined in Table E.2 and available round-robin test data shown in Figure E.1. For EXP07 repeatability assessment there are 9 datasets, 7 based on tests done at UL and 2 based on tests done at PG&E. It should be noted that with AFS and NEEA4, three tests were performed based on EXP07 at UL and the test unit was once removed and reinstalled in test rooms in between those three tests. Similarly, with NEEA7, the first two tests (P2 and P3) were performed based on EXP07 at UL and then one final test #11 with NEEA7B was performed at UL. So, the repeatability evaluation results using all UL tests for AFS, NEEA4, and NEEA7(B) are somewhat hybrid between true repeatability and reproducibility results as the unit was once removed and re-installed in between repeated tests. As can be seen in Table E.3, two different data sets are defined for UL EXP07 testing of both units NEEA4 and NEEA7(B) in order to assess the impact of reinstallation on repeatability. In contrast, the EXP07 PG&E repeatability results for the same unit models are based on tests that were repeated without test unit reinstallation. The datasets NEEA4-UL-E-3R and NEEA7(B)-UL-E-3R include all three sets of tests for these units, whereas the NEEA4-UL-E-2W and NEEA7-UL-E-2W datasets only include the two back-to-back tests, excluding the other test where the test unit was reinstalled in the test facility. However, for AFS EXP07 repeatability evaluation, only one dataset, AFS-UL-E-3R, was defined using all three EXP07 tests. This is because overall repeatability was good for all three tests, indicating that reinstallation at UL did not cause significant issues. For AHRI 210/240 repeatability evaluation, there are 5 datasets across two labs. In addition, for EXP07 and AHRI 210/240 reproducibility assessment, there are two datasets for each using all the tests at UL and PG&E for NEEA4 and NEEA7(B). There are 5 tests in each dataset for EXP07 reproducibility assessment and 4 tests in each dataset for AHRI 210/240 reproducibility evaluation.

Table E.3 Datasets for EXP07 and AHRI 210/240 repeatability and reproducibility analysis based on different set of tests in two labs for multiple units

	Dataset	Test Unit	Test Lab [Tests in Dataset]	Test Methodology
Repeatability	AFS-UL-E-3R	AFS	UL [4 & 6 & 12]	EXP07:2019
	NRCan8-UL-E-2W	NRCan8	UL [1 & 2]	EXP07:2019
	NRCan10-UL-E-2W	NRCan10	UL [3 & 4]	EXP07:2019
	NEEA4-UL-E-3R	NEEA4	UL [P1 & 8 & 10]	EXP07:2019
	NEEA4-UL-E-2W	NEEA4	UL [8 & 10]	EXP07:2019
	NEEA4-PGE-E-2W	NEEA4	PG&E [6 & 8]	EXP07:2019
	NEEA7(B)-UL-E-3R	NEEA7 & NEEA7B	UL [P2 & P3 & 11]	EXP07:2019
	NEEA7-UL-E-2W	NEEA7	UL [P2 & P3]	EXP07:2019
	NEEA7B-PGE-E-2W	NEEA7B	PG&E [2 & 4]	EXP07:2019
	AFS-UL-A-2W	AFS	UL [3 & 5]	AHRI 210/240-2023
	NEEA4-UL-A-2W	NEEA4	UL [7 & 9]	AHRI 210/240-2023
	NEEA4-PGE-A-2W	NEEA4	PG&E [5 & 7]	AHRI 210/240-2023
	NEEA7-UL-A-2W	NEEA7	UL [1 & 2]	AHRI 210/240-2023
	NEEA7B-PGE-A-2W	NEEA7B	PG&E [1 & 3]	AHRI 210/240-2023
Reproducibility	NEEA4-UL&PGE-E-5R	NEEA4	UL [P1 & 8 & 10], PG&E [6 & 8]	EXP07:2019
	NEEA7(B)-UL&PGE-E-5R	NEEA7 & NEEA7B	UL [P2 & P3 & 11], PG&E [2 & 4]	EXP07:2019
	NEEA4-UL&PGE-A-4R	NEEA4	UL [7 & 9], PG&E [5 & 7]	AHRI 210/240-2023
	NEEA7(B)-UL&PGE-A-4R	NEEA7 & NEEA7B	UL [1 & 2], PG&E [1 & 3]	AHRI 210/240-2023

As per EXP07, a test unit's performance is measured in two sets of cooling test intervals, dry coil and humid coil, and two sets of heating test intervals, continental and marine, which represent different climate types. Then, measured performance in different test intervals is propagated through a temperature bin method to estimate the cooling and heating seasonal coefficient of performance (SCOP) for different climate zones. Table E.4 shows different cooling and heating climate zones as per EXP07 along with corresponding test type results used in SCOP estimation.

Table E.4 EXP07 cooling and heating climate zones and test type used in SCOP estimation

Cooling Climate Zone	**Very Cold**	**Cold/Dry**	**Cold/ Humid**	**Marine**	**Mixed**	**Hot/ Humid**	**Hot/Dry**	
Test Type	Humid	Dry	Humid	Dry	Humid		Dry	
Heating Climate Zone	**Subarctic**	**Very Cold**	**Cold/Dry**	**Cold/ Humid**	**Marine**	**Mixed**	**Hot/ Humid**	**Hot/Dry**
Test Type	Continental				Marine	Continental		

Figure E.2 presents the repeatability assessment based on 7 different datasets (Table E.3) for various test units tested at UL showing the minimum, maximum, and mean of two statistical parameters across different climate zones for estimated cooling and heating SCOP based on repeated tests. ΔSCOP (Max-Min) is the difference in the minimum and maximum estimated SCOP of different tests for each climate zone described as the percentage of the tests' mean SCOP. It represents the maximum variation in performance between different tests. SCOP STD is the standard deviation of SCOP based on repeated tests shown as a percentage of the average SCOP of multiple tests. This represents the overall variation in estimated seasonal performance normalized to the number of repeated tests, since the number of repeated tests varied in different datasets for different units as shown in Figure E.1 and Table E.3.

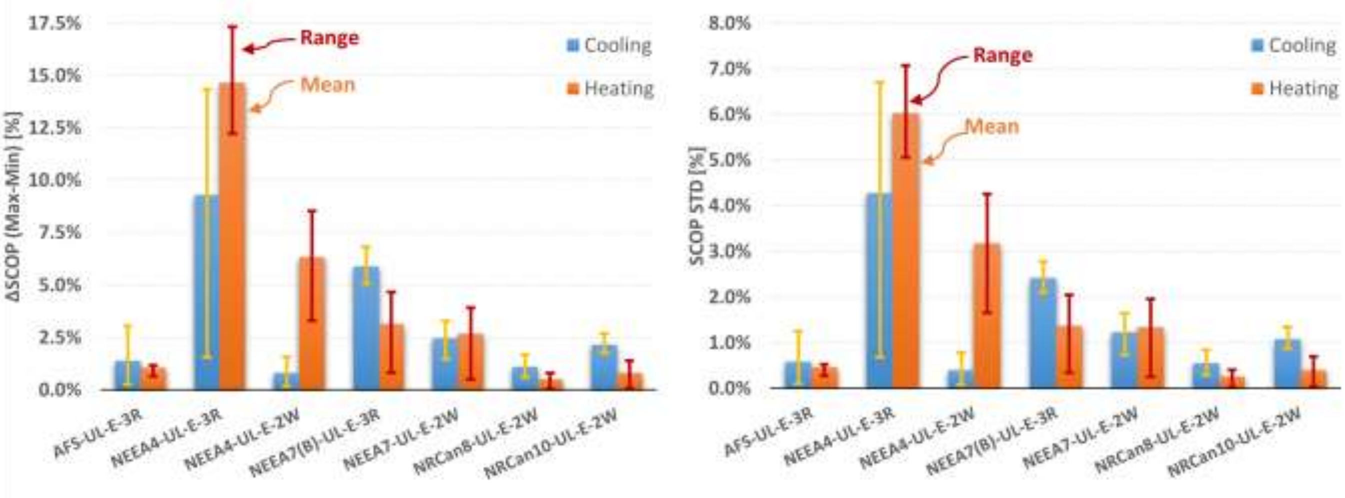


Figure E.2 EXP07 cooling and heating SCOP repeatability assessment summary for different datasets from UL with mean and range of statistical parameters across different climate zones

Overall, good repeatability was observed in cooling and heating estimated SCOP of AFS, NRCan8, and NRCan10 test units at UL as shown with results based on datasets AFS-UL-E-3R, NRCan8-UL-E-2W and NRCan10-UL-E-2W. Compared to these three datasets, a relatively larger variation was observed in measured performance among three tests in dataset NEEA7(B)-UL-E-3R with the NEEA7 and NEEA7B unit, where reasonable repeatability was observed with somewhat better results in heating mode compared to cooling mode tests. Furthermore, good repeatability is achieved in both heating and cooling mode when considering the NEEA7-UL-E-2W dataset. That dataset only considers back-to-back tests (P2 and P3) at UL with NEEA7 without including the last test #11 with NEEA7B at UL that was performed after reinstallation about two years later. NEEA4 test results for dataset NEEA4-UL-E-3R showed somewhat poor repeatability, which was mostly due to significantly different measured performance in test P1 that was performed more than a year earlier compared to the other two sets of tests, #8 and #10. Without taking test P1 into account in the NEEA4 repeatability evaluation in dataset NEEA4-UL-E-2W, the test unit shows good repeatability in cooling mode and reasonable repeatability in heating mode. The difference in unit performance during test P1 compared to the other two tests is mostly attributable to differences in test unit dynamic response under the same test conditions and unit dynamic behavior captured during the convergence period in some test intervals as can be seen in section 4.2 of the report. Figure 21, Figure 23 to Figure 25, Figure 26, and Figure 28 to Figure 31 show differences in test unit dynamic response and behavior captured during the convergence period for cooling tests with NEEA4 at UL. Similar differences for heating tests with NEEA4 at UL are shown in Figure 35 and Figure 37 to Figure 45. It is important to note that test P1 with NEEA4 at UL was performed around 16 months before the other two tests, #8 and #10, and the test unit was taken out of the test rooms after test P1 and re-installed. So, in addition to the variability in test unit integrated controls response, differences in NEEA4 dynamic response at the same test conditions could be due to differences in test setup, charge, testing approach, etc. It should also be noted that tests P2 and P3 with NEEA7 were also performed around the same time as test P1 with NEEA4 and around 24 months prior to the 3rd test (#11) with NEEA7B. However, unit NEEA7(B) (NEEA7 & NEEA7B) based on dataset NEEA7(B)-UL-E-3R showed better repeatability than NEEA4 with dataset NEEA4-UL-E-3R based on all three sets of tests at UL.

Figure E.3 shows the EXP07 repeatability assessment summary results based on datasets of two test units at PG&E. Good overall repeatability was observed for both test units, NEEA4 and NEEA7B, in cooling as well as heating mode. In comparison to UL results, PG&E test results for both units showed better repeatability. This could be because, unlike at UL, the PG&E tests were performed back-to-back with no unit re-installation in between. As a result, there was no variability associated with the test setup, refrigerant charge, instrumentation, and other similar factors.

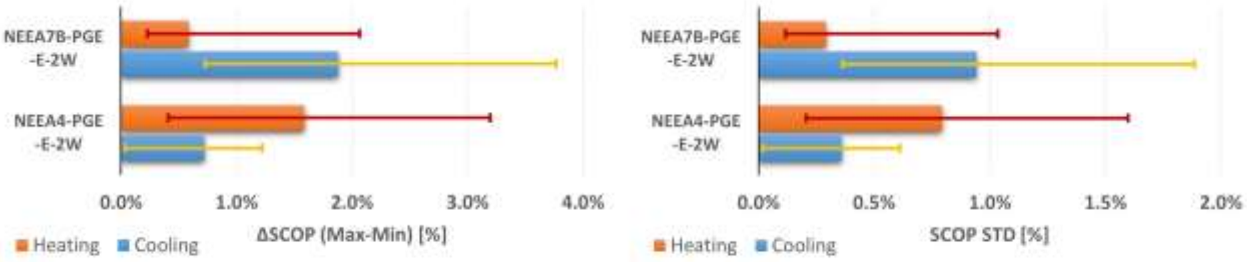


Figure E.3 EXP07 cooling and heating SCOP repeatability assessment summary for two datasets from PG&E with mean and range of statistical parameters across different climate zones

Figure E.4 shows summary results of the EXP07 reproducibility assessment for two test units, based on results from multiple tests at UL and PG&E. In NEEA4 cooling tests at PG&E, full-load tests were erroneously performed in some test intervals where load-based tests should have been conducted, due to a possible issue with the testing approach implementation. This resulted in some large differences in cooling SCOP between the two labs, and because of this testing approach issue, comparing NEEA4 cooling SCOP between the two labs does not provide relevant conclusions regarding EXP07 reproducibility assessment and thus results for this case are not shown in the plots. Furthermore, heating test results for NEEA4 showed poor reproducibility. Compared to NEEA4, NEEA7(B) demonstrated better reproducibility in both cooling and heating results. Overall good reproducibility was observed for NEEA7(B) heating test results, whereas the test unit showed relatively poorer reproducibility for cooling test results.

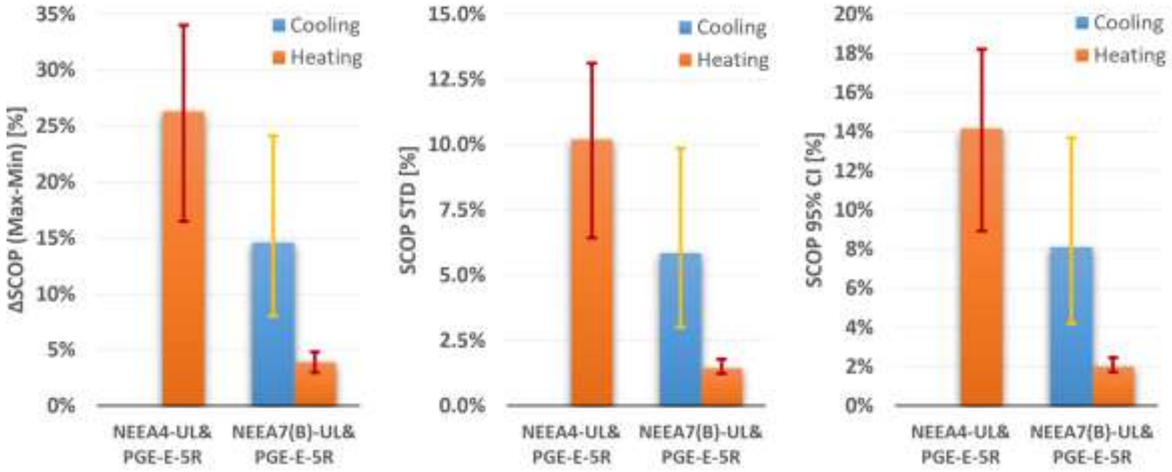


Figure E.4 EXP07 cooling and heating SCOP reproducibility assessment summary for two test units for testing at UL and PG&E with mean and range of statistical parameters across different climate zones

Table E.5 provides a categorization of different factors that contribute to the overall repeatability and reproducibility of the EXP07 test results along with a description of their influences and variability.

Table E.5 Categorization of variables contributing to EXP07 repeatability and reproducibility

No.	Factor	Repeatability Issues	Reproducibility Issues
1	**Unit under Test (UUT) Installation/Setup -** includes differences in refrigerant charge, duct static pressure setup, thermostat offset setup, type & installation of instrumentation	Relevant when unit is re-installed between tests, especially when installations occur many months apart, when different personnel are involved, and if interpretation of the standard setup changes	An inherent issue for any test standard, but load-based testing has additional installation requirements that are less familiar to personnel and that are still evolving
2	**Environmental Chamber Characteristics -** includes differences in air flow and temperature distribution, responsiveness of reconditioning controls	Potentially an issue if a UUT were re-installed within a different chamber between tests or at a different location within the same chamber; also an issue if chamber controls do not track indoor or outdoor setpoints well due to UUT dynamic behavior	Differences in air flow and temperature distribution could impact dynamic response of the thermostat; in some cases, chamber controls may not track indoor or outdoor setpoints well due to UUT dynamic behavior
3	**Interpretation and Implementation of Method of Test -** includes convergence criteria, transition to full-load tests	Especially relevant if different personal involved in implementation and/or when repeatability tests are separated over long duration. Also, there are some shortcomings with the current convergence criteria.	An inherent issue for any test standard, but load-based testing has additional requirements that are less familiar to personnel and that are still evolving
4	**Dynamic Behavior of UUT -** covers variations in dynamic	Lack of repeatability due to inconsistent control behavior is	Lack of reproducibility due to inconsistent control behavior

	behavior of UUT due to variation in control behavior that may result from adaptive learning and limited time period available for testing	an inherent issue for some equipment that probably needs to be addressed by each manufacturer	is an inherent issue for some equipment that probably needs to be addressed by each manufacturer
5	**UUT Replacement** - due to failure or performance degradation	Performance varies due to manufacturing tolerances and/or firmware differences	Performance varies due to manufacturing tolerances and/or firmware differences

Table E.6 provides a qualitative assessment of the cooling and heating mode repeatability results for each unit tested in the two labs along with the possible factors as per Table E.5 contributing to the observed differences between repeated tests. The factors considered to be the most important in contributing to observed large differences between different tests are described. As mentioned previously, the UL repeatability results for AFS, NEEA4, and NEEA7(B) based on all three repeated tests at UL, as shown with datasets AFS-UL-E-3R, NEEA4-UL-E-3R, and NEEA7(B)-UL-E-3R are somewhat between true repeatability and reproducibility results because the units were re-installed once in between repeated tests. This could be one of the primary factors for relatively poor repeatability with the NEEA4-UL-E-3R and NEEA7(B)-UL-E-3R datasets at UL compared to other datasets. For both of these units, when only considering two sets of back-to-back tests at UL, in datasets NEEA4-UL-E-2W and NEEA7-UL-E-2W, repeatability improves for both cooling and heating modes as shown in Table E.6.

Table E.6 EXP07 repeatability assessment summary for different datasets in two labs along with key factors contributing to observed differences

Dataset	Repeatability Assessment		Important Factors
	Cooling	Heating	
AFS-UL-E-3R	Good	Good	**1** - unit was re-installed after being in storage for a period of time and then retested.
NEEA4-UL-E-3R	Poor	Poor	**1, 3, 4** - unit was re-installed after a period of time and then re-tested; convergence criteria implementation; UUT had inconsistent dynamic response and also difference in measured airflow at same test conditions. Refer to Figure 21, Figure 23 to Figure 25, Figure 26, Figure 28 to Figure 31, Figure 35, Figure 37 to Figure 44, and Figure 45 in section 4.2 for examples.
NEEA4-UL-E-2W	Good	Fair	**3, 4 -** differences in dynamic response captured with convergence criteria in different tests; UUT had inconsistent dynamic response at same test conditions. Refer to the same figures as provided in above cell to see some of differences in NEEA4 test #8 and #10 which are included in this dataset.
NEEA7(B)-UL-E-3R	Fair	Good	**1, 3, 4, 5** – unit re-installed after period of time and re-tested; UUT was replaced due to some issues while setting up for testing at PG&E lab; UUT had inconsistent dynamic response; also had differences in dynamic response captured with convergence criteria. Refer to Figure 68, Figure 70 to Figure 73, Figure 74, Figure 76,

			Figure 77, Figure 80, Figure 82 to Figure 84, and Figure 85 in section 4.4.
NEEA7-UL-E-2W	Good	Good	**3, 4** - some differences in dynamic response captured with convergence criteria in different tests; UUT had inconsistent dynamic response at same test conditions. Refer to the same figures as provided for NEEA7(B)-UL-E-3R to see some of differences in NEEA7 tests P2 and P3 which are included in this dataset.
NRCan8-UL-E-2W	Good	Good	**-**
NRCan10-UL-E-2W	Good	Good	**-**
NEEA4-PGE-E-2W	Good	Good	**-**
NEEA7B-PGE-E-2W	Good	Good	**-**

Similarly, Table E.7 shows overall reproducibility assessment results for two units along with possible factors contributing to observed differences with a description of the most important factors.

Table E.7 EXP07 reproducibility assessment summary for two datasets along with key factors contributing to observed differences

Dataset	**Reproducibility Assessment**		**Important Factors**
	Cooling	**Heating**	
NEEA4-UL&PGE-E-5R	-	Poor	**1, 2, 3, 4** - differences associated with the facilities and UUT installation/setup including thermostat offset settings and measured airflow; UUT had inconsistent dynamic response; differences in convergence criteria implementation and dynamic response captured during convergence period; also difference in transition to full-load. Refer to Figure 167, Figure 169Figure 170 to Figure 173, Figure 174, Figure 176, Figure 177, Figure 181, Figure 183 to Figure 187, and Figure 188 in section 5.2.
NEEA7(B)-UL&PGE-E-5R	Fair	Good	**1, 2, 3, 4, 5** - differences associated with the facilities and UUT installation/setup; measured airflow; UUT had inconsistent dynamic response; differences in convergence criteria implementation and dynamic response captured during convergence period; also UUT was replaced in between tests. Refer to Figure 135, Figure 137 to Figure 141, Figure 142, Figure 144, Figure 145, Figure 149, Figure 151 to Figure 158, Figure 159, and Figure 161 to Figure 163 in section 5.1.

It should be clear from the results of Table E.6 and Table E.7 that performance results are much less consistent after equipment has been re-installed in either the same or a different laboratory than if the equipment is retested with the same installation. Some of the important factors that affect the reproducibility include issues related to the installation and setup of the unit for testing (e.g., instrumentation, refrigerant charge, duct static pressure, etc.), different interpretations and implementations of the draft standard, and different characteristics of the environmental chambers. Although these factors are also relevant for the current standard (AHRI 210/240), their impact on the results for load-based testing might be more significant because the draft standard is new and more complicated and the dynamic behavior of the unit with its integrated controls may be sensitive to some installation effects. In addition to the effect of installation and implementation issues on test unit dynamic behavior, variations in test unit dynamic response at the same test conditions could be due to adaptive/learning behavior of the embedded controller and the limited time available for testing. Some of the differences in implementation will likely be resolved as the standard matures and personnel become more familiar with its application. Some of the issues with differences in facilities could be addressed by facility and test equipment improvements. For example, utilizing a thermostat apparatus to provide a standardized environment for the thermostat could reduce differences that are caused by non-uniform airflow and temperatures within environmental chambers. However, differences related to changes in test unit controller behavior would be challenging to address with the test standard without dramatically increasing the testing time. In fact, one of the merits of the load-based testing approach is that it can capture the impacts and sensitivities of test unit performance to dynamic responses of its embedded controller. If a test unit controller performs inconsistently with load-based testing in the laboratory, then it is likely to perform similarly in the field and it is important to capture these effects in a standard method of test. This could be an incentive for manufacturers to develop controllers with more consistent and predictable behavior. With that being said, it would be a good idea to further investigate how best to test and rate units that have inconsistent controller behavior using load-based testing in a repeatable and representative fashion for making the standard more inclusive.

There were likely some differences in implementation of EXP07 convergence criteria that led to some significant performance differences between tests carried out in different labs. For instance, differences in test unit operation mode and dynamic behavior that were captured during periods that were deemed to be converged were likely due to different interpretations of the EXP07 convergence criteria and some limitations of the current convergence criteria. These convergence criteria issues occurred when there was a combination of different operating modes that occurred at a given test condition, such as a sequential combination of unit on/off cycling and defrost, or when there was an irregular on/off cycling pattern with a mixture of short and regular cycles. Examples of these issues can be seen in NEEA4 heating test results time series plots from two labs in section 5.2.2. These issues are discussed in detail in the report with recommended improvements which could be used as a reference to update the EXP07 convergence criteria.

Another specific issue identified through testing was that indoor temperatures could converge to significantly different values, most notably when comparing test results between labs. This could primarily be related to differences in the thermostat offset settings at the start of testing that are part of implementation of the EXP07 setup procedures as seen in Figure 183 to Figure 185 in section 5.2.2 for NEEA4 heating test results. Another EXP07 implementation issue that led to disparities between laboratories is related to interpretation of the decision of when to switch to full-load testing as seen in Figure 172 and Figure 173 showing cooling test results of NEEA4 in two labs in section 5.2.1. Prematurely switching to full-load testing can have a significant impact on results at relatively high-load conditions that are important in determining seasonal efficiencies.

Some of the possible sources of differences in test unit performance that are related to test setup (e.g., charge, thermal mass, instrumentation location, sensor dynamic responses, air flow measurement, test unit fan settings) and test facility control (e.g., room conditions, external static pressure) could be addressed by improving test setup requirements and/or corrections for the components that have a significant effect on dynamic performance measurements, such as code-tester thermal mass, instrumentation location, sensor response, etc. Also, more appropriate test condition and operating tolerances should be defined to limit the effect of variations in test facility control on overall performance measurement. Both of these enhancements, however, will necessitate additional research and examination because most of the currently used test methods are for steady-state tests rather than dynamic load-based tests. Also, issues with external static pressure control, airflow measurement, and test fan settings that led to differences in the two labs should be further investigated with the goal of updating the current EXP07 testing approach to have better reproducibility.

The repeatability and reproducibility of AHRI 210/240 is also subject to some of the same test setup issues that affect load-based testing. However, the setup and implementation of this standard is more familiar to personnel and generally less complicated. Also, the tests are steady state and therefore additional uncertainties related to dynamic measurements are not relevant. In order to evaluate repeatability and reproducibility of AHRI 210/240 for comparison with load-based testing, the standard was implemented for the test units and laboratories identified in Figure E.1. The measured performance for different test intervals was propagated through a temperature bin method to estimate cooling and heating seasonal performance metrics, SEER (Seasonal Energy Efficiency Ratio) and HSPF (Heating Seasonal Performance Factor). Figure E.5 shows the average SEER and HSPF for each unit based on corresponding datasets as defined in Table E.3 from UL and PG&E labs along with a 95% confidence interval (CI) for AHRI 210/240 repeatability and reproducibility evaluation. Table E.8 shows a summary of AHRI 210/240 repeatability and reproductivity assessment with the average value of SEER and HSPF based on different tests along with the percentage difference in maximum and minimum value, standard deviation (STD), and 95% CI as statistical parameters to show variability in estimated seasonal performance.

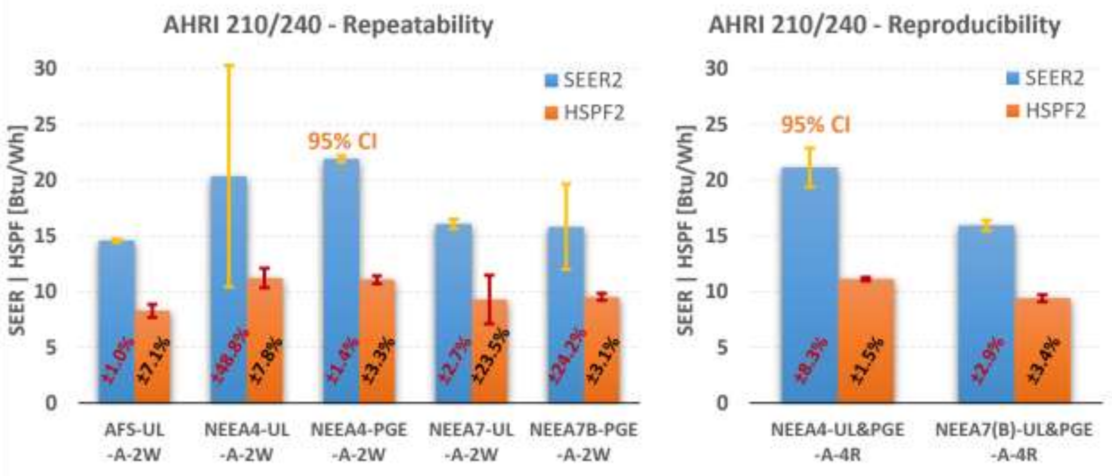


Figure E.5 AHRI 210/240 repeatability and reproducibility summary with average seasonal performance metric and 95% CI based on different tests

Table E.8 AHRI 210/240 repeatability and reproducibility assessment summary results

AHRI 201/240		Repeatability					Reproducibility	
Metric	**Dataset**	**AFS-UL-A-2W**	**NEEA4-UL-A-2W**	**NEEA4-PGE-A-2W**	**NEEA7-UL-A-2W**	**NEEA7B-PGE-A-2W**	**NEEA4-UL&PGE-A-4R**	**NEEA7(B)-UL&PGE-A-4R**
SEER2	Mean	14.56	20.34	21.91	16.06	15.80	21.12	15.93
	∆(Max - Min) [%]	0.2%	7.7%	0.2%	0.4%	3.8%	11.2%	3.8%
	STD [%]	±0.1%	±3.8%	±0.1%	±0.2%	±1.9%	±4.5%	±1.6%
	95% CI [%]	±1.0%	±48.8%	±1.4%	±2.7%	±24.2%	±8.3%	±2.9%
HSPF2	Mean	8.27	11.20	11.06	9.28	9.52	11.13	9.40
	∆(Max - Min) [%]	1.1%	1.2%	0.5%	3.7%	0.5%	2.2%	4.6%
	STD [%]	±0.6%	±0.6%	±0.3%	±1.8%	±0.2%	±0.8%	±1.8%
	95% CI [%]	±7.1%	±7.8%	±3.3%	±23.5%	±3.1%	±1.5%	±3.4%

AFS unit showed good repeatability at UL in two repeated tests as shown with the AFS-UL-A-2W dataset results. Good repeatability was also observed for the NEEA4 unit at UL and PG&E as shown with datasets NEEA4-UL-A-2W and NEEA4-PGE-A-2W, except for cooling test results at UL in dataset NEEA4-UL-A-2W. In cooling test intervals at UL with NEEA4, relatively poor performance was measured in test #7 compared to the test #9 in dataset NEEA4-UL-A-2W. Further, for the NEEA4 unit, higher SEER was estimated based on PG&E tests in dataset NEEA4-PGE-A-2W compared to UL tests in dataset NEEA4-UL-A-2W, resulting in relatively poor reproducibility in cooling test results based on dataset NEEA4-UL&PGE-A-4R. However, the NEEA4 unit had slightly better HSPF that was estimated based on PG&E tests in dataset NEEA4-PGE-A-2W compared to UL tests in dataset NEEA4-UL-A-2W, and overall good reproducibility was noted in heating test results based on dataset NEEA4-UL&PGE-A-4R. NEEA7 test results showed good repeatability in the UL tests in dataset NEEA7-UL-A-2W as well as NEEA7B unit in PG&E lab tests in dataset NEEA7B-PGE-A-2W. At UL, relatively better repeatability was noted in cooling test results compared to heating in dataset NEEA7-UL-A-2W, whereas the opposite was observed in PG&E test results in dataset NEEA7B-PGE-A-2W. Between the two labs, overall good reproducibility was observed for cooling as well as heating results based on dataset NEEA7(B)-UL&PGE-A-4W. The differences in the two labs are possibly due to differences in the test setup, charge, instrumentation, airflow rate measurement (Figure 207 and Figure 212), external static pressure control, and measurement uncertainties.

Figure E.6 and Figure E.7 compares the repeatability and reproducibility of EXP07 and AHRI 210/240 in terms of percentage STD applied to cooling and heating seasonal performance metrics using the test data from UL and PG&E for NEEA4 and NEEA7(B) (NEEA7 & NEEA7B). STD is used as a statistical parameter rather than 95% CI because a different number of tests were used in the AHRI 210/240 and EXP07 assessments for some cases. For EXP07 results, a mean and range of cooling and heating SCOP STD as observed across different climate zones is shown, whereas single SEER and HSPF STD values are presented for AHRI results in a single climate zone. The results for reproducibility will first be discussed, which can be assessed by viewing the results for all tests of each unit across both labs in Figure E.6 and Figure E.7. The datasets utilized for EXP07 and AHRI 210/240 reproducibility assessment for NEEA4 and NEEA7(B) can be seen in Table E.3. For both NEEA4 and NEEA7(B), significantly better reproducibility is observed in AHRI 210/240 results compared to EXP07, except for NEEA7(B) heating results, where comparable variations in seasonal performance metrics were observed for both testing approaches. Some possible reasons for the poor reproducibility of EXP07 have previously been discussed.

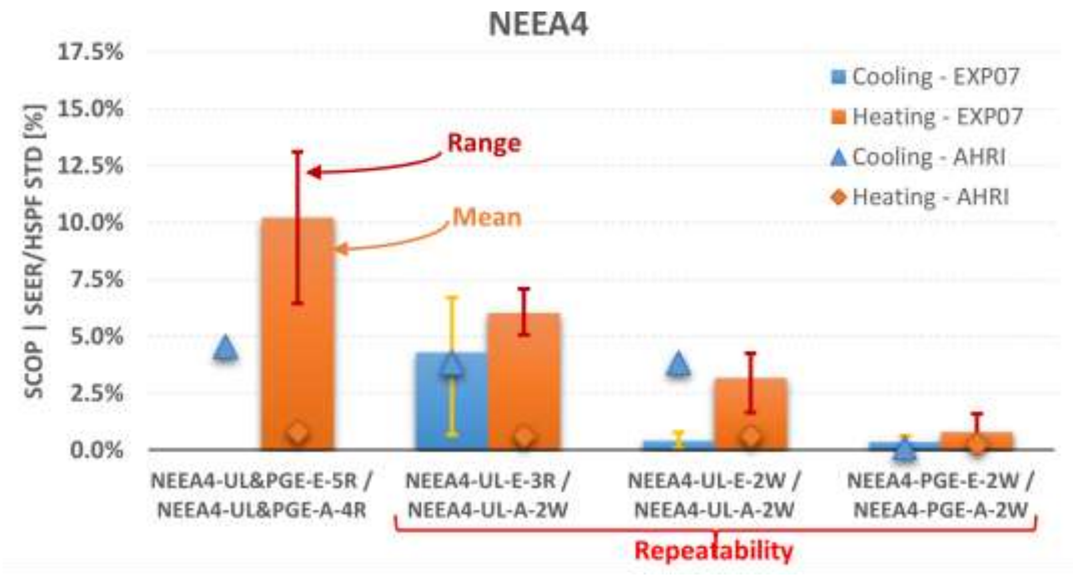


Figure E.6 NEEA4 EXP07 and AHRI 210/240 seasonal performance metrics STD comparison for reproducibility and repeatability assessment. EXP07 SCOP STD mean and range across different climate zones and AHRI SEER / HSPF STD for single climate zone

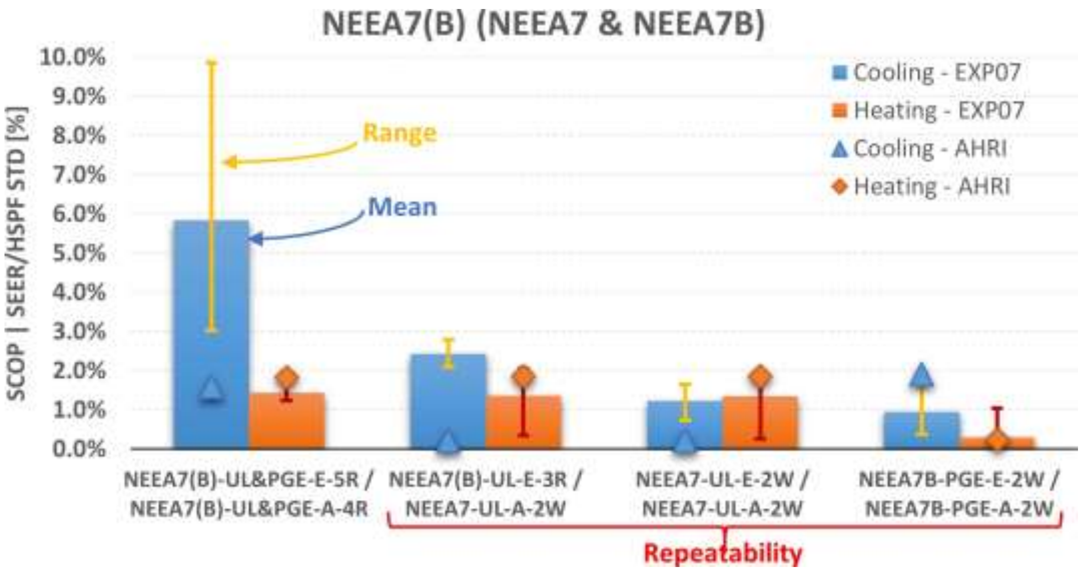


Figure E.7 NEEA7(B) EXP07 and AHRI 210/240 seasonal performance metrics STD comparison for reproducibility and repeatability assessment. EXP07 SCOP STD mean and range across different climate zones and AHRI SEER / HSPF STD for single climate zone

Repeatability can be assessed by viewing results in Figure E.6 and Figure E.7 that were determined from test results at each individual lab (UL and PG&E). Note that the AHRI 210/240 repeatability tests for units NEEA4 and NEEA7(B) involved only two tests at each lab that were performed sequentially without having to reinstall the unit as defined in datasets NEEA4-UL-A-2W, NEEA4-PGE-A-2W, NEEA7-UL-A-2W, and NEEA7B-PGE-A-2W. The EXP07 tests at UL involved three tests where only two of the tests were back-to-back without reinstallation as defined in datasets NEEA4-UL-E-3R and NEEA7(B)-UL-E-3R. In order to provide fairer repeatability comparisons, results for UL EXP07 that only include back-to-back tests without reinstallation are also shown (NEEA4-UL-E-2W and NEEA7-UL-E-2W datasets). Overall, the repeatability of both EXP07 and AHRI 210/240 are good and comparable in both cooling and heating when not considering data where the unit was not reinstalled. In fact, there are some examples where the EXP07 repeatability was better than for AHRI 210/240 (e.g., cooling tests for NEEA4 at UL). If data is included for units that were reinstalled at UL, then the repeatability of AHRI 210/240 would generally be better than that for EXP07, especially for the NEEA4 heating results. However, it is felt that these comparisons are less relevant for repeatability because they contain several factors related to reproducibility.

This report presents several suggestions for improving the draft EXP07:19 standard that resulted from analysis of the repeatability and reproducibility data along with interactions with the project participants and stakeholders as outlined in section 6. These are mostly related to convergence criteria, test unit setup, transition to full-load testing, unit control settings, airflow measurement, thermostat offsets, and tolerances associated with dynamic measurements. Many of the improvements may have already been addressed in the most recent updates to EXP07:19 that have led to the latest version (EXP07:22) that is scheduled to be published in August 2022. For future work in the near term, it is suggested that the raw test data that was obtained in this project be re-analyzed to assess whether revisions to the convergence criteria address some of the noted issues and improve overall repeatability and reproducibility. For the longer term, it is strongly recommended that a similar study be carried to assess the overall impact of updates that are incorporated in EXP07:22. A second study will benefit from experiences gained during the initial study. However, a second study should include more test units across more than just two test laboratories and be focused on reproducibility. Although there was sufficient data in the initial study to gain an understanding of the repeatability of the EXP07 testing, the data were not sufficient to draw clear conclusions on its reproducibility. Ideally, reproducibility of test results for both EXP07 and AHRI 210/240 should be considered for all test units across all test labs included in a follow up project.

A key question that needs to be addressed in future work is: what is an acceptable tolerance in seasonal performance metric repeatability and reproducibility for load-based testing per EXP07? An initial step that is necessary to answer this question involves carrying out a detailed uncertainty analysis for measured performance that is due to dynamic measurements that are subject to both instrumentation uncertainties and dynamic behavior associated with the dynamic testing approach, the testing facilities, and operating tolerances. This should lead to the definition of minimum tolerances. However, the actual tolerances would need to be larger to account for human factors that lead to differences in test installation and setup in different labs. These tolerances can only be defined through testing data obtained across multiple labs when applying the updated EXP07:22 draft standard. Given the dynamic nature of tests in the load-based testing approach, these tolerances will most likely be larger than what is expected for AHRI 210/240. So, there is undoubtedly a tradeoff when comparing AHRI 210/240 and EXP07 between reproducibility and field representativeness of the test approaches. It is too soon to fully understand this tradeoff before additional work is carried out. In particular, it is important to analyze AHRI 210/240 and EXP07 test results with a goal of identifying the primary sources of differences between the two approaches along with their contributions to overall estimated seasonal performance differences. It is also important to consider paths towards even better testing and rating approaches for the future. In particular, there could be tremendous value in defining test approaches that could determine performance maps for heat pumps and air conditioners that could be integrated into simulation tools for estimating more accurate and climate-specific seasonal performance ratings. Testing of equipment with integrated controls could be important in developing performance maps for this purpose. Future work to develop procedures for testing and performance mapping that minimize test requirements while enhancing reproducibility should be a priority.

Table of Contents

1 Introduction

This report presents the repeatability and reproducibility assessment of two testing and rating methodologies for residential heat pumps, one based on a load-based testing approach as per CSA EXP07:19 and the other based on AHRI 210/240-2023. Round robin testing was done with five different heat pumps of varied sizes and types in two labs, UL and PG&E, for both test procedures. This study provides a quantitative and qualitative analysis of test results with these systems from a repeatability and reproducibility evaluation perspective of both testing methodologies along with a root cause analysis of the observed performance variations in different tests. Here, first, the testing approach and test conditions are described along with a brief description of the test units. Following that, an overview of the data analysis approach is outlined. Then EXP07 repeatability and reproducibility evaluation analysis results are presented for each test unit, which includes the comparison of cooling and heating seasonal coefficient of performance (SCOP) for different tests along with a quantitative and qualitative comparison of test unit performance and dynamic behavior in different test intervals along with any possible issues identified with a test interval. For some test intervals, test unit dynamic performance with time is also presented to explain the possible causes of observed performance differences at the same test conditions. Then similarly repeatability and reproducibility assessment analysis results are presented for AHRI 210/240 along with a discussion on possible reasons for observed differences. Also, a summary of the repeatability and reproducibility assessment for both test methodologies is provided along with some recommendations to further improve the load-based testing methodology. Finally, conclusions and possible future work are discussed.

2 Testing Approach and Test Conditions

For repeatability and reproducibility analysis of CSA EXP07:19 (CSA, 2019) and AHRI 210/240-2023 (AHRI, 2020), the approach is to test five different units multiple times at the same lab (UL) and then test two of those units at a different lab (PG&E) and compare the test results within the same lab and in between different labs for each unit. Table 1 provides a brief description of the test units and their rated performance as per AHRI 210/240-2017 (AHRI, 2017).

Table 1. Test units' description and rated performance

Test Unit	**AFS**	**NEEA4**	**NEEA7 & NEEA7B**	**NRCan8**	**NRCan10**
Description	5-ton split-system ducted (single-stage)	1-ton min-split ductless (variable-speed)	3-ton split-system ducted (variable-speed)	1.5-ton min-split ductless (variable-speed)	2-ton split-system ducted (variable-speed)
Q_C [Btu/h] @ 95°F	55000	10900	33000	14000	24200
EER @ 95°F	13	12.5	12.5	13	14.5
SEER	16	20	17.8	21.6	19
Q_H [Btu/h] @ 47°F	55000	13600	38000	18000	24000
HSPF (Region IV)	8.5	12	11	11.7	10.5
Q_H [Btu/h] @ 17°F	33600	8800	29000	12100	-

Figure 1 shows the progression of testing as per EXP07 and AHRI 210/240 for different test units in the two labs. Each box named either CSA EXP07 and AHRI 210/240-2023 represents one complete set of cooling and heating test intervals based on the corresponding test methodology in the test lab highlighted on the left-hand side. In this report "test interval" refers to a single test at a particular test condition and "test" refers to the entire set of test intervals as per the test methodology associated with either EXP07 or AHRI 210/240. For example, as per EXP07, one test consists of 5 cooling dry coil, 4 cooling humid coil, 6 heating continental, and 4 heating marine test intervals as shown in Table 4 and Table 5. In Figure 1, the number in the box on the right side adjacent to each test shows the corresponding test name or number. This test name or number in combination with the test unit and test lab provide a unique identifier for each set of test intervals data. Note that while referring to test names with "PREV" in their name, a short form "P" is generally used in the text in place of "PREV", for example, "PREV1" and "P1" refer to the same test.

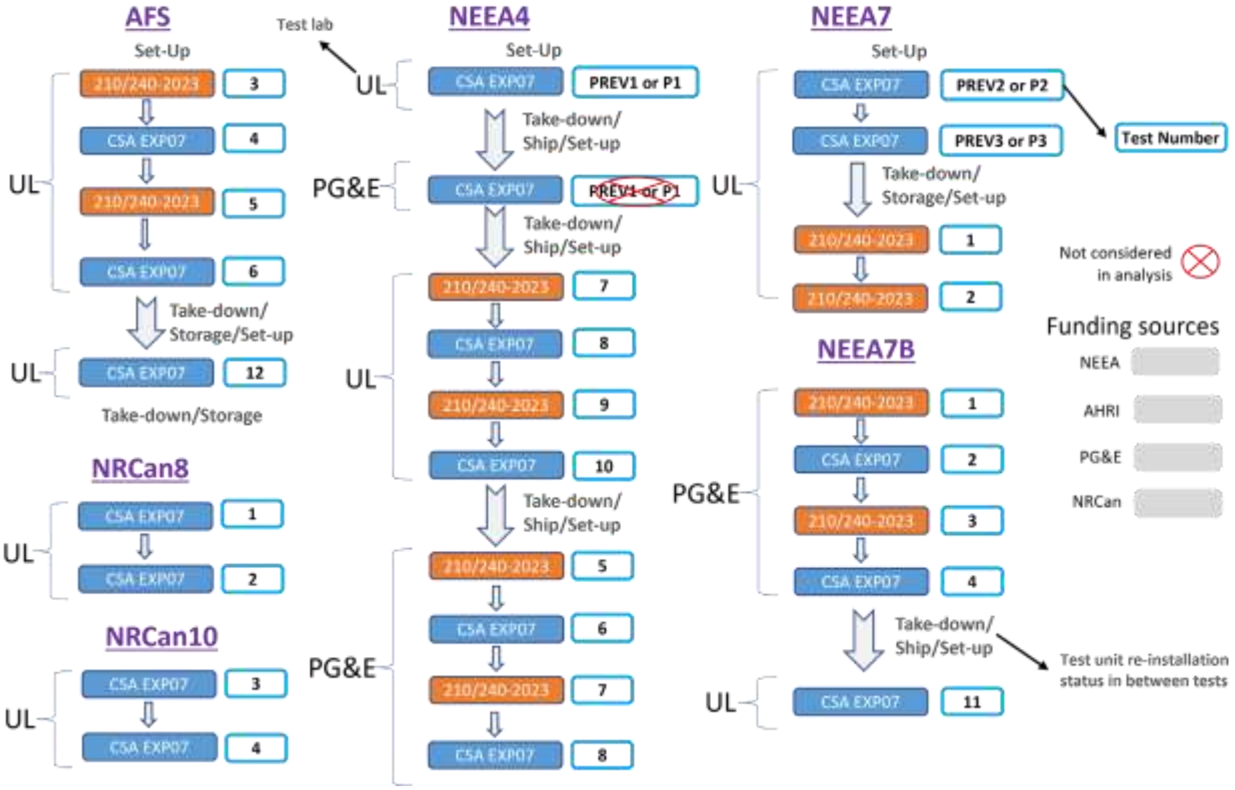


Figure 1. Round robin tests sequence for repeatability and reproducibility analysis of CSA EXP07:19 and AHRI 210/240-2023

Figure 1 also highlights whether a test unit was taken out of test rooms and re-installed in between different tests either in same or different lab. Further for the NEEA4 unit, test P1 data from the PG&E lab was not considered in the analysis, as at that time load-based testing was being set-up at the PG&E lab and confidence in the consistency of collected data is low. It can be seen that three units, AFS, NRCan8, and NRCan10, were only tested at UL, and the performance of NRCan8 and NRCan10 was only measured based on EXP07. As a result, the tests for these three units were only utilized to examine the repeatability of the test methodologies. Since the other two test unit models, NEEA4 and NEEA7 & NEEA7B, were tested in both labs based on both test approaches, the results were utilized to assess repeatability and reproducibility for both methods. One thing to note regarding NEEA7 is that after the initial phase of UL testing when it was shipped to PG&E, it was discovered during setup that the unit was not operating properly. As a result, a different unit but the same model, NEEA7B, was used for PG&E tests and one final test at UL. It should be noted that even though test unit was changed from NEEA7 to NEEA7B in the middle of the test sequence, the results of these two different units were still compared for overall repeatability and reproducibility assessment of EXP07 and AHRI 210/240 as the same model was utilized in testing. Further, here NEEA7(B) name is used to indicate that the analysis results being discussed include the test results of both test units NEEA7 and NEEA7B.

For CSA EXP07 and AHRI 210/240 repeatability assessment, each test unit's performance across different tests from the same lab is compared, whereas, for reproducibility assessment, NEEA4 and NEEA7(B) (NEEA7 & NEEA7B) performance from different tests across two labs are compared. For repeatability and reproducibility evaluation, different datasets are defined based on the nomenclature approach outlined in Table 2. Each dataset name is defined based on the test unit, test lab, test methodology, number of tests in the dataset, and whether the test unit was reinstalled or not in between different tests in the dataset, either in in the same lab or different labs.

Table 2. Dataset nomenclature approach

Test Unit –	**Test Lab** –	**Test Methodology** –	**No. of Tests in Dataset** |	**Unit Reinstallation Status between Tests**
AFS **NRCan8** **NRCan10** **NEEA4** **NEEA7** **NEEA7B** **NEEA7(B):** NEEA7 & NEEA7B	**UL** **PGE:** PG&E **UL&PGE:** UL and PG&E	**E:** EXP07-2019 **A:** AHRI 210/240-2023	**2** **3** **4** **5** (Individual tests included in dataset are provided in Table 3)	**R:** Test unit reinstalled at least once in between different tests in the dataset, either in the same lab or in different labs **W:** Tests were performed without test unit reinstallation in between different tests in the dataset

Table 3 shows the defined datasets for EXP07 and AHRI 210/240 repeatability and reproducibility evaluation based on the nomenclature outlined in Table 2 and available round-robin test data shown in Figure 1. For EXP07 repeatability assessment there are 9 datasets, 7 based on tests done at UL and 2 based on tests done at PG&E. It should be noted that with AFS and NEEA4, three tests were performed based on EXP07 at UL and the test unit was once removed and reinstalled in test rooms in between those three tests. Similarly, with NEEA7, the first two tests (P2 and P3) were performed based on EXP07 at UL and then one final test #11 with NEEA7B was performed at UL. So, the repeatability evaluation results using all UL tests for AFS, NEEA4, and NEEA7(B) are somewhat hybrid between true repeatability and reproducibility results as the unit was once removed and re-installed in between repeated tests. As can be seen in Table 3, two different data sets are defined for UL EXP07 testing of both units NEEA4 and NEEA7(B) in order to assess the impact of reinstallation on repeatability. In contrast, the EXP07 PG&E repeatability results for the same unit models are based on tests that were repeated without test unit reinstallation. The datasets NEEA4-UL-E-3R and NEEA7(B)-UL-E-3R include all three sets of tests for these units, whereas the NEEA4-UL-E-2W and NEEA7-UL-E-2W datasets only include the two back-to-back tests, excluding the other test where the test unit was reinstalled in the test facility. However, for AFS

EXP07 repeatability evaluation, only one dataset, AFS-UL-E-3R, was defined using all three EXP07 tests. This is because overall repeatability was good for all three tests, indicating that reinstallation at UL did not cause significant issues. For AHRI 210/240 repeatability evaluation, there are 5 datasets across two labs. In addition, for EXP07 and AHRI 210/240 reproducibility assessment, there are two datasets for each using all the tests at UL and PG&E for NEEA4 and NEEA7(B). There are 5 tests in each dataset for EXP07 reproducibility assessment and 4 tests in each dataset for AHRI 210/240 reproducibility evaluation.

Table 3. Datasets for EXP07 and AHRI 210/240 repeatability and reproducibility analysis based on different set of tests in two labs for multiple units

	Dataset	Test Unit	Test Lab [Tests in Dataset]	Test Methodology
Repeatability	AFS-UL-E-3R	AFS	UL [4 & 6 & 12]	EXP07:2019
	NRCan8-UL-E-2W	NRCan8	UL [1 & 2]	EXP07:2019
	NRCan10-UL-E-2W	NRCan10	UL [3 & 4]	EXP07:2019
	NEEA4-UL-E-3R	NEEA4	UL [P1 & 8 & 10]	EXP07:2019
	NEEA4-UL-E-2W	NEEA4	UL [8 & 10]	EXP07:2019
	NEEA4-PGE-E-2W	NEEA4	PG&E [6 & 8]	EXP07:2019
	NEEA7(B)-UL-E-3R	NEEA7 & NEEA7B	UL [P2 & P3 & 11]	EXP07:2019
	NEEA7-UL-E-2W	NEEA7	UL [P2 & P3]	EXP07:2019
	NEEA7B-PGE-E-2W	NEEA7B	PG&E [2 & 4]	EXP07:2019
	AFS-UL-A-2W	AFS	UL [3 & 5]	AHRI 210/240-2023
	NEEA4-UL-A-2W	NEEA4	UL [7 & 9]	AHRI 210/240-2023
	NEEA4-PGE-A-2W	NEEA4	PG&E [5 & 7]	AHRI 210/240-2023
	NEEA7-UL-A-2W	NEEA7	UL [1 & 2]	AHRI 210/240-2023
	NEEA7B-PGE-A-2W	NEEA7B	PG&E [1 & 3]	AHRI 210/240-2023
Reproducibility	NEEA4-UL&PGE-E-5R	NEEA4	UL [P1 & 8 & 10], PG&E [6 & 8]	EXP07:2019
	NEEA7(B)-UL&PGE-E-5R	NEEA7 & NEEA7B	UL [P2 & P3 & 11], PG&E [2 & 4]	EXP07:2019
	NEEA4-UL&PGE-A-4R	NEEA4	UL [7 & 9], PG&E [5 & 7]	AHRI 210/240-2023
	NEEA7(B)-UL&PGE-A-4R	NEEA7 & NEEA7B	UL [1 & 2], PG&E [1 & 3]	AHRI 210/240-2023

Table 4 and Table 5 show the cooling and heating test conditions for different test intervals based on CSA EXP07:19. The primary test outcome from each test interval is the unit under test (UUT) coefficient of performance (COP), which is then propagated through a temperature bin method approach to estimate the cooling and heating seasonal coefficient of performance (SCOP) for different climate zones. Table 6 and Table 7 show the different climate zones, test interval data used in SCOP estimation for that climate zone, and bin hour fractions for cooling and heating, respectively, from CSA EXP07:19. Further details on the load-based test approach can be found in CSA EXP07:19 (CSA, 2019). It should be noted that, at PG&E, test unit performance was not measured in test interval at -10°F outdoor temperature due to test facility limitations.

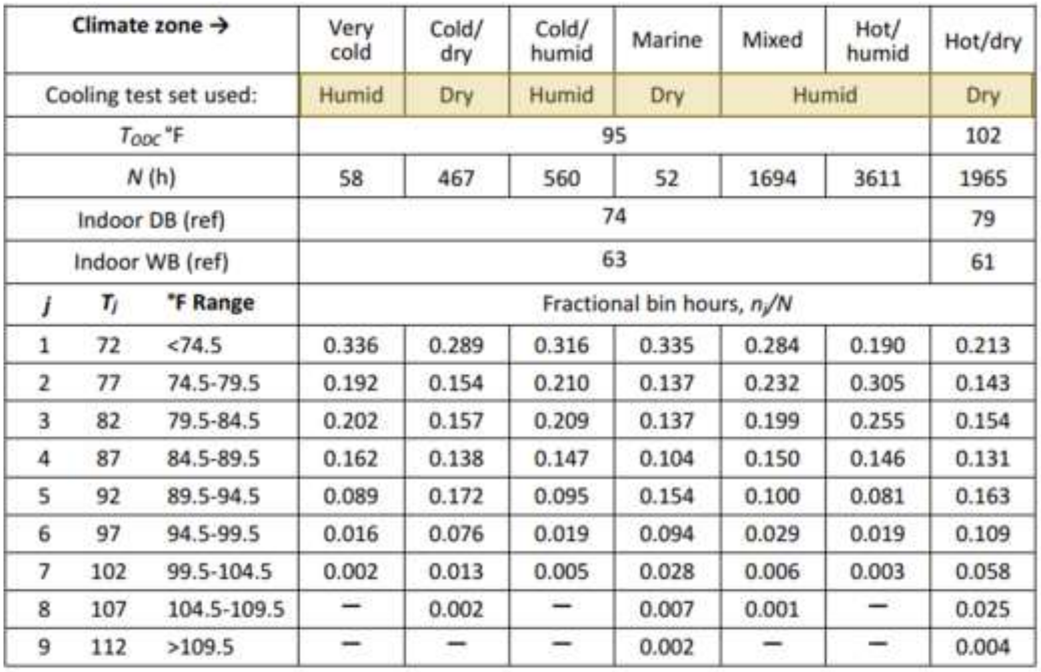
Table 4. CSA EXP07:19 Cooling Test Conditions (CSA, 2019)

Test Name	Humid test conditions			Dry test conditions		
	Outdoor temperature [°F]	Indoor temperature [°F]		Outdoor temperature [°F]	Indoor temperature [°F]	
	Dry Bulb	Dry Bulb	Wet Bulb	Dry Bulb	Dry Bulb	Wet Bulb
CA	N/A	74	63	113	79	56 (maximum)
CB	104			104		
CC	95			95		
CD	86			86		
CE	77			77		

Table 5. CSA EXP07:19 Heating Test Conditions (CSA, 2019)

Test Name	Continental outdoor conditions		Marine outdoor conditions		Indoor conditions	
	Dry-bulb temperature [°F]	Wet-bulb temperature [°F]	Dry-bulb temperature [°F]	Wet-bulb temperature [°F]	Dry-bulb temperature [°F]	Wet-bulb temperature [°F]
HA	-10	-11.4	N/A	N/A	70	60 (maximum)
HB	5	4				
HC	17	14.5	17	15.5		
HD	34	31	34	32		
HE	47	41	47	45		
HF	54	45	54	49		

Table 6. Bin hour fractions and test results used for calculating cooling seasonal coefficient of performance ($SCOP_C$) (CSA, 2019)



Climate zone →			Very cold	Cold/ dry	Cold/ humid	Marine	Mixed	Hot/ humid	Hot/dry
Cooling test set used:			Humid	Dry	Humid	Dry	Humid		Dry
T_{ODC} °F			95						102
N (h)			58	467	560	52	1694	3611	1965
Indoor DB (ref)			74						79
Indoor WB (ref)			63						61
j	T_j	**°F Range**	Fractional bin hours, n_j/N						
1	72	<74.5	0.336	0.289	0.316	0.335	0.284	0.190	0.213
2	77	74.5-79.5	0.192	0.154	0.210	0.137	0.232	0.305	0.143
3	82	79.5-84.5	0.202	0.157	0.209	0.137	0.199	0.255	0.154
4	87	84.5-89.5	0.162	0.138	0.147	0.104	0.150	0.146	0.131
5	92	89.5-94.5	0.089	0.172	0.095	0.154	0.100	0.081	0.163
6	97	94.5-99.5	0.016	0.076	0.019	0.094	0.029	0.019	0.109
7	102	99.5-104.5	0.002	0.013	0.005	0.028	0.006	0.003	0.058
8	107	104.5-109.5	—	0.002	—	0.007	0.001	—	0.025
9	112	>109.5	—	—	—	0.002	—	—	0.004

Table 7. Bin hour fractions and test results used for calculating heating seasonal coefficient of performance ($SCOP_H$) (CSA, 2019)

Climate zone →			Sub-arctic	Very cold	Cold/ Dry	Cold/ Humid	Marine	Mixed	Hot/ Humid	Hot/ Dry
Heating test set used:			Standard				Marine	Standard		
T_{ODH} °F			-40	-10	4	-1	30	16	33	26
N			7758	6340	5017	4808	4630	3299	1199	2162
Indoor DB (ref)			70							
j	***T_j***	**°F Range**	**Fractional bin hours, *n_j/N***							
1	–23	< –20.5	0.043	0.004	-	-	-	-	-	-
2	–18	–20.5 - –15.5	0.024	0.006	0.001	0.002	-	-	-	-
3	–13	–15.5 - –10.5	0.028	0.009	0.002	0.003	-	-	-	-
4	-8	–10.5 - –5.5	0.036	0.015	0.004	0.007	-	-	-	-
5	–3	–5.5–0.5	0.038	0.019	0.005	0.010	-	0.001	-	-
6	2	–0.5 - 4.5	0.050	0.028	0.010	0.020	-	0.003	-	0.001
7	7	4.5–9.5	0.050	0.035	0.014	0.029	-	0.007	-	0.002
8	12	9.5–14.5	0.051	0.045	0.025	0.041	-	0.015	0.001	0.004
9	17	14.5–19.5	0.062	0.061	0.047	0.062	0.002	0.033	0.005	0.011
10	22	19.5–24.5	0.057	0.067	0.064	0.071	0.005	0.046	0.015	0.022
11	27	24.5–29.5	0.089	0.102	0.120	0.116	0.019	0.094	0.049	0.054
12	32	29.5–34.5	0.123	0.136	0.162	0.150	0.061	0.137	0.098	0.093
13	37	34.5–39.5	0.114	0.145	0.171	0.151	0.147	0.171	0.165	0.145
14	42	39.5–44.5	0.076	0.108	0.125	0.104	0.199	0.147	0.173	0.161
15	47	44.5–49.5	0.082	0.108	0.123	0.114	0.282	0.161	0.217	0.218
16	52	49.5–54.5	0.055	0.077	0.079	0.078	0.205	0.118	0.171	0.175
17	57	> 54.5	0.022	0.034	0.047	0.041	0.079	0.065	0.105	0.114

Table 8 and Table 9 show the cooling and heating test conditions, respectively, for single stage and variable speed equipment as per AHRI 210/240. In the last column, the green highlighted cells show the test intervals in which corresponding type test units' performance was measured in two labs. Further, it should also be noted that at the PG&E lab, NEEA7(B) performance was not measured in the optional heating test interval H42. Further, the primary output from AHRI 210/240 test intervals is measured cooling or heating rate and power consumption at different ambient conditions with proprietary control settings for compressor speeds and airflow rates. Also, there is an optional test interval to measure the cyclic degradation coefficient, which was only conducted for the AFS unit here. For the other two units, the default value as per AHRI 210/240 was utilized in seasonal performance estimation. Then measured performance at different test conditions is propagated through a temperature bin method to estimate cooling and heating seasonal performance metrics, SEER (Seasonal Energy Efficiency Ratio) and HSPF (Heating Seasonal Performance Factor). Further information regarding testing and seasonal performance estimation approach can be found in AHRI 210/240-2023 (AHRI, 2020).

Table 8. AHRI 210/240-2023 Cooling Test Conditions (AHRI, 2020)

Test Name	Air Entering Outdoor Unit (ºF)		Air Entering Indoor Unit (ºF)		Compressor Speed	Indoor Airflow	Test Requirement	
	Dry Bulb	Wet Bulb	Dry Bulb	Wet Bulb			Single Stage	Variable Speed
A2	95	75	80	67	Full$_C$	Full$_C$	R	R
B2	82	65	80	67	Full$_C$	Full$_C$	R	R
B1	82	65	80	67	Low$_C$	Low$_C$		R
C2	82	58	80	57	Full$_C$	Full$_C$	O	
D2	82	58	80	57	Full$_C$	Full$_C$	O	
Ev	87	69	80	67	Int$_C$	Int$_C$		R
F1	67	53.5	80	67	Low$_C$	Low$_C$		R

Table 9. AHRI 210/240-2023 Heating Test Conditions (AHRI, 2020)

Test Name	Air Entering Outdoor Unit (ºF)		Air Entering Indoor Unit (ºF)		Compressor Speed	Indoor Airflow	Test Requirement	
	Dry Bulb	Wet Bulb	Dry Bulb	Wet Bulb			Single Stage	Variable Speed
H01	62	56.5	70	60	Low$_H$	Low$_H$		R
H12	47	43	70	60	Full$_H$	Full$_H$	R	O
H11	47	43	70	60	Low$_H$	Low$_H$		R
H1C	47	43	70	60	Full$_H$	Full$_H$	O	O
H1N	47	43	70	60	Nom$_H$	Nom$_H$		R
H22	35	33	70	60	Full$_H$	Full$_H$	R	O
H2v	35	33	70	60	Int$_H$	Int$_H$		R
H32	17	15	70	60	Full$_H$	Full$_H$	R	R
H42	5	3	70	60	Full$_H$	Full$_H$	O	O

3 Data Analysis Approach

In this study repeatability and reproducibility assessment for CSA EXP07:19 and AHRI 210/240-2023 is performed based on test results of multiple test units in two labs, UL and PG&E as shown in Figure 1, based on the defined datasets in Table 3. For a test methodology repeatability evaluation, first, test intervals performance results are compared based on the repeated tests for a test unit within the same lab followed by the comparison of seasonal performance metrics to assess the overall performance variability in different tests. For reproducibility assessment, test intervals and overall estimated seasonal performance metric results were compared for each test unit based on multiple tests in two labs.

In this report, EXP07 repeatability and reproducibility assessment results are presented first. For each test unit, for a quantitative and qualitative comparison of performance in different tests at the same test conditions, measured COP and observed test unit dynamic behavior during convergence are compared for each test interval. Table 10 shows the nomenclature for different prominent operation modes of UUT observed during load-based tests based on CSA EXP07:19. Further, for root cause analysis of the observed performance differences in different tests under the same test conditions, possible issues with a test interval are also identified which are divided into 3 main categories, as shown in Table 11. These are related to convergence criteria, inconsistent unit dynamic response, and testing approach. It should be noted that possible issues noted with a test interval do not necessarily contribute to the variations observed in test unit performance at the same test condition. In some cases, it was just noted to highlight the possible further areas of improvement for the load-based testing approach. To augment the root cause analysis, some key test intervals plots illustrating the test unit performance with time are also presented along with a discussion on some of the main reasons for observed performance variations. In addition to each test interval performance comparison, estimated cooling and heating SCOP were also compared between different tests across different climate zones to show the overall variation in seasonal performance metrics.

Table 10. UUT Operation Mode Nomenclature

Abbr.	Unit Operation Mode Behavior
OO	ON/OFF Cycling, Regular (regular, consistent ON/OFF pattern)
OOH	ON/OFF Cycling Hybrid, Irregular (irregular, inconsistent ON/OFF pattern)
VS	Variable Steady (the compressor loads steadily and stably; does not cycle OFF)
VH	Variable Hybrid (the compressor loads and unloads aggressively (>20% step); does not cycle OFF)
DF	Defrost
FL	Full Load Test
NC	Not Converged (unit performance during test interval did not converge)

Table 11. Possible Issues with a Test Interval

Abbr.	Possible Issue Flag for a test Interval
CC	Convergence Criteria (either criterion itself as outlined in EXP07 or its implementation)
UR	Unit Response (inconsistent unit dynamic response)
TA	Testing Approach (either the testing approach itself as outlined in EXP07 or its implementation)

To quantify the variation in measured performance in each test interval and overall SCOP, mainly three different statistical parameters were utilized. One is the difference in the minimum and maximum value of a performance metric between different tests shown as a percentage of the mean value among different tests to illustrate the maximum variation in a performance metric. Second is the 95% confidence interval (CI) for the average performance metric estimated based on the results from different tests. A confidence interval is estimated for both cooling and heating SCOP's in different climate zones and COP's at different test intervals. As there are only a limited number of tests (samples), a student's tt-distribution is assumed for the test unit average performance, and a 95% confidence is calculated accordingly based on the number of samples (nn) available as per the equations below.

$$CCCCnnCCCCCCCCnnCCCC\ IInnttCCIIIIIII = X\text{�} - tt_{0.95}\frac{\sigma\sigma_{ss}}{\sqrt{nn}} \tag{1}$$

$$wwhCCIICC;\ X\text{�} = \frac{\sum xx_{ii}}{nn};\ \sigma\sigma_{ss} = \text{�}\frac{\sum(xx_{ii} - X\text{�})^2}{nn - 1};\ tt_{0.95}\text{: } 4.303\ (nn = 3), 12.706\ (nn = 2), \ldots \tag{2}$$

As *t*-distribution confidence interval is also a strong function of the number of samples, which might lead to a larger confidence interval with a smaller number of samples (tests) for a unit even though the overall variation in measured performance is relatively small. Using confidence intervals to compare the performance variation results between two units with a different number of tests can lead to erroneous conclusions. For this, population standard deviation (STD) is a better statistical parameter which is basically root mean squared error compared to mean as shown in the equation below.

$$SSSSSS = \overline{\text{�}\frac{\sum(xx_{ii} - X\text{�})^2}{nn}} \tag{3}$$

In plots, both 95% CI and STD are mostly presented as a percentage of the mean performance metric. A similar data analysis approach was used to assess AHRI 210/240-2023 repeatability and reproducibility, by first comparing the test interval performance along with a root cause analysis of observed differences and then finally comparing the overall seasonal performance metrics between different tests.

In the subsections below, first, EXP07 repeatability evaluation results are presented for different test units in two labs, followed by EXP07 reproducibility assessment results, and then an overall summary of EXP07 repeatability and reproducibility evaluation along with some recommendations to further improve the load-based testing approach. Following that, AHRI 210/240 repeatability and reproducibility evaluation results are presented.

4 EXP07 Repeatability Assessment

In this section, CSA EXP07:19 repeatability assessment is presented based on the repeated test results of 5 different units at two different labs, UL and PG&E, utilizing different datasets defined in Table 3. All 5 units were tested at UL; however, only 2 of those (NEEA4 and NEEA7B) were tested at PG&E. In the subsections below, a repeatability assessment of each unit in the corresponding lab is presented for cooling and heating mode. First, measured COP and observed test unit behavior (Table 10) during convergence at each test interval are compared along with any possible issue flag (Table 11) for a quantitative and qualitative comparison of measured performance in repeated tests. For further root cause analysis, some key test intervals plots showing the test unit performance with time are also shown. Then, the seasonal coefficient of performance (SCOP) estimated based on measured performance in repeated tests is compared to assess the overall repeatability. Additionally, cooling load-based test result plots for some units depicting the variation of test unit performance, indoor temperature, and humidity are presented in Appendix A, mainly to present the humidity response during load-based tests.

4.1 AFS (UL)

Here, repeatability assessment results of test unit AFS, a 5-ton fixed-speed heat pump system, tested at UL lab are presented, first for cooling mode and then for heating mode for dataset AFS-UL-E-3R.

4.1.1 Cooling Mode

Figure 2 shows the comparison of cooling dry coil test intervals COP and unit operation mode behavior during convergence at different outdoor temperatures between 3 repeated tests at UL. This single-stage unit cycled on/off at low ambient temperatures (77°F, 86°F, and 95°F), operated in variable-speed mode at 104°F, and failed to meet the building load at 113°F where a full-load test was performed. The test unit showed a similar operation mode in repeated tests at the same test conditions. Figure 3 shows a quantitative comparison of COP for the same test intervals between repeated tests. The left plot shows COP at different ambient test conditions along with the percentage difference of each test interval COP compared to the mean COP of three tests at the same test conditions. The right plot shows the mean value of COP with standard deviation (STD), a statistical parameter to show the variation in COP between different tests at the same test conditions. In the right plot, the difference between the maximum and minimum COP measured in different tests at the same test conditions is also shown as a percentage to mean value (ΔMax%) on the right vertical axis.

Overall, the test unit showed good repeatability in cooling dry coil test intervals with maximum difference within 3.1% and STD within ±1.3% except at 113°F outdoor condition test interval (CA). The main difference in UUT performance for the full-load test at 113°F test interval results from relatively lower COP in test #4 compared to tests #6 and #12 with a maximum difference among the three of around 13.1%. This relatively lower performance in this test interval for test #4 seems to have resulted from higher static pressure compared to other tests as can be seen in Figure 4 and Figure 5. There was a step-change in the static pressure that occurred 15 minutes into this test that was due to the shutdown of the code tester blower. This resulted in the test unit blower changing the airflow because of its constant torque motor and correspondingly lower cooling rate and higher power consumption. This seems like an issue with test facility control rather than load-based testing methodology and ideally, this test interval should not be considered in the repeatability assessment of the test methodology.

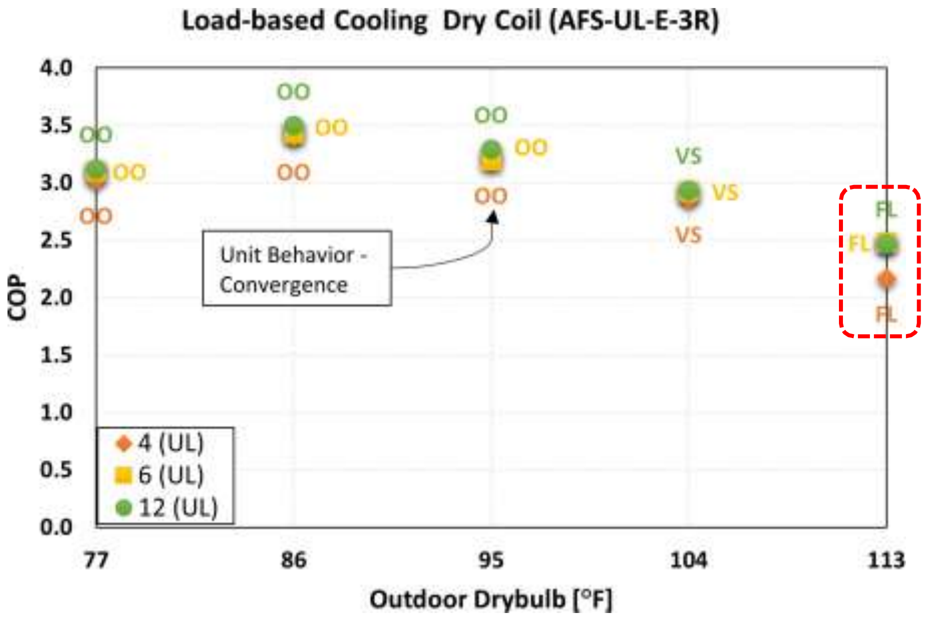


Figure 2. AFS (UL) cooling dry coil test intervals COP for different tests with unit behavior during convergence

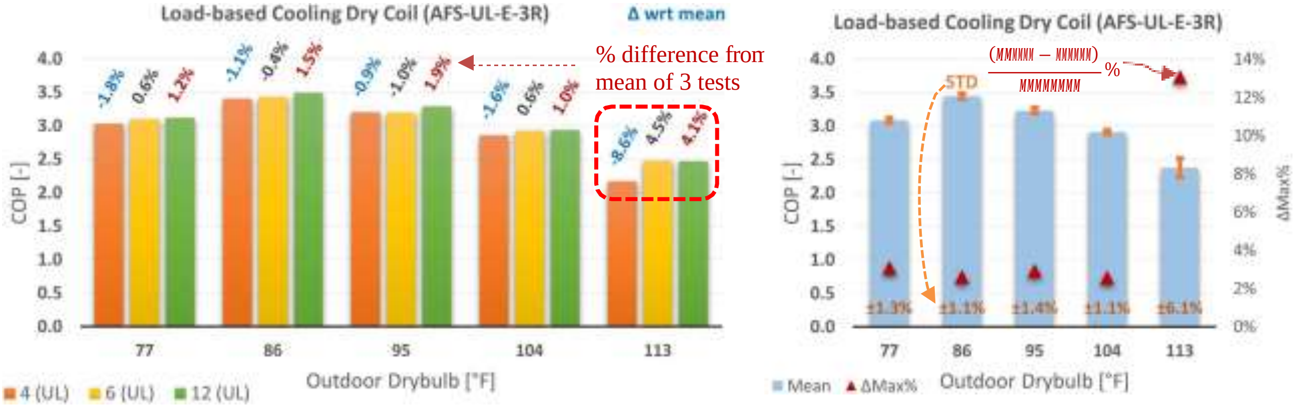


Figure 3. AFS (UL) cooling dry coil test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Another interesting thing observed was performance convergence in variable-speed operation mode (i.e., building load met without unit cycling on/off) during 104°F test interval (CB) for this fixed-speed unit. Figure 6 shows the test unit performance with time for this test interval during test #12. UUT cycled off at the beginning of the test and then convergence was found later in 40 minutes window during compressor *on* period as per EXP07 convergence criteria. However, it seems like the unit would have cycled off again if ran it for a longer duration and that would have resulted in on/off behavior. A further investigation and discussion are required on whether the current convergence criteria capture representative overall performance for this kind of behavior or need to modify it.

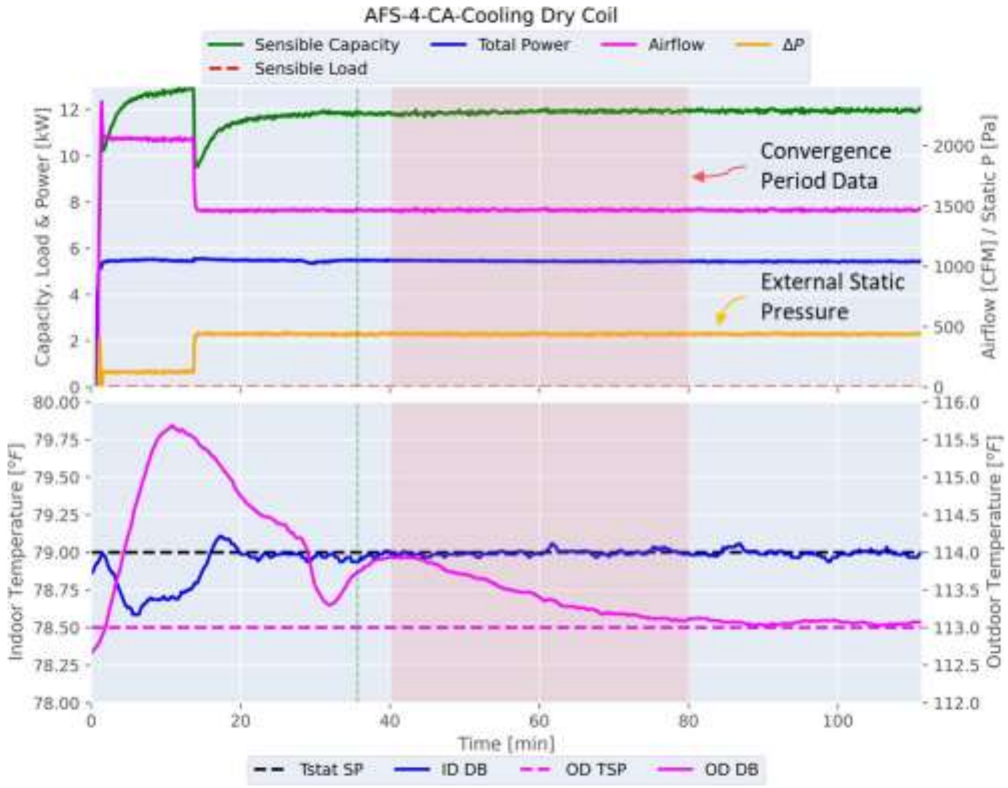


Figure 4. AFS-4 (UL) performance and temperature variations for cooling dry coil test interval CA

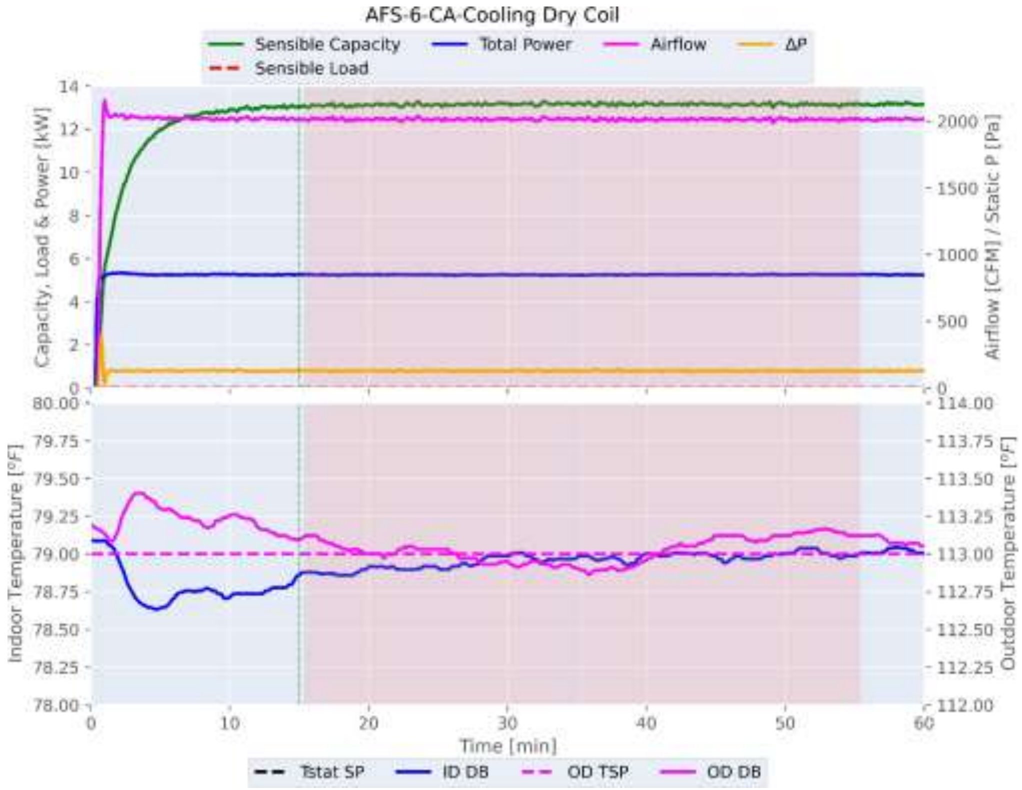


Figure 5. AFS-6 (UL) performance and temperature variations for cooling dry coil test interval CA

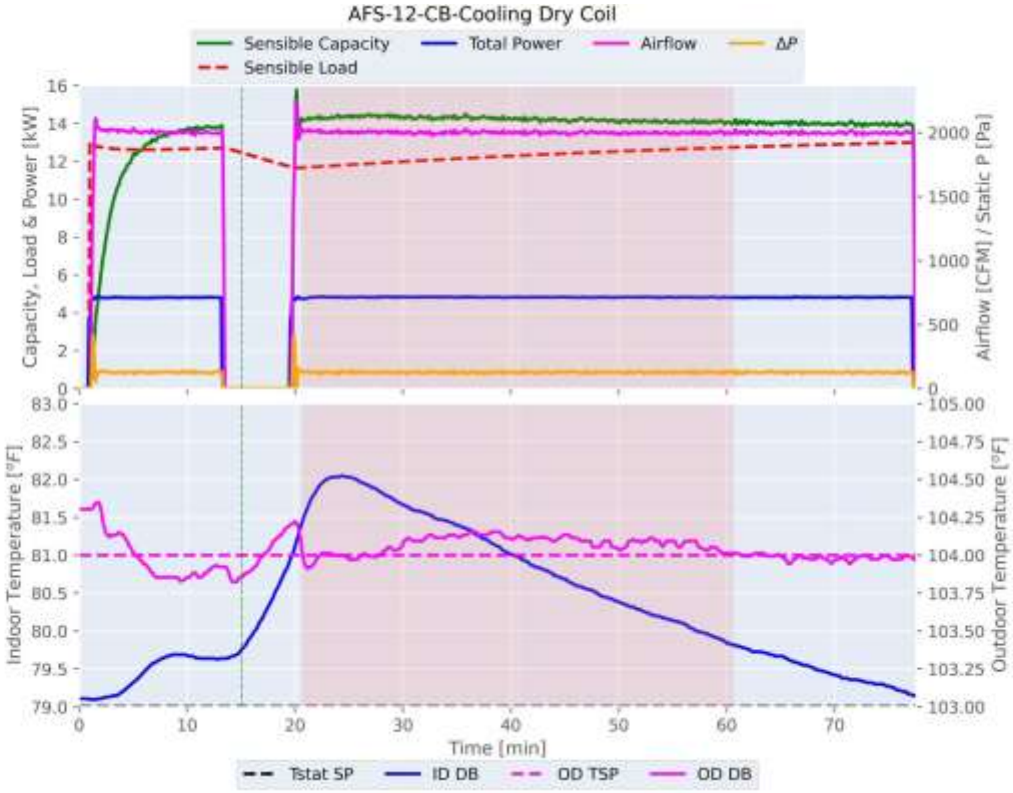


Figure 6. AFS-12 (UL) performance and temperature variations for cooling dry coil test interval CB (104℉)

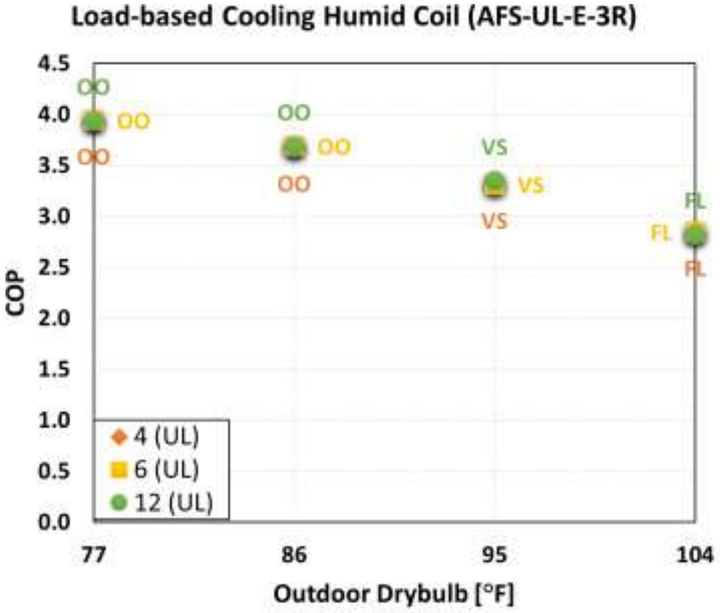


Figure 7. AFS (UL) cooling humid coil test intervals COP for different tests with unit behavior during convergence

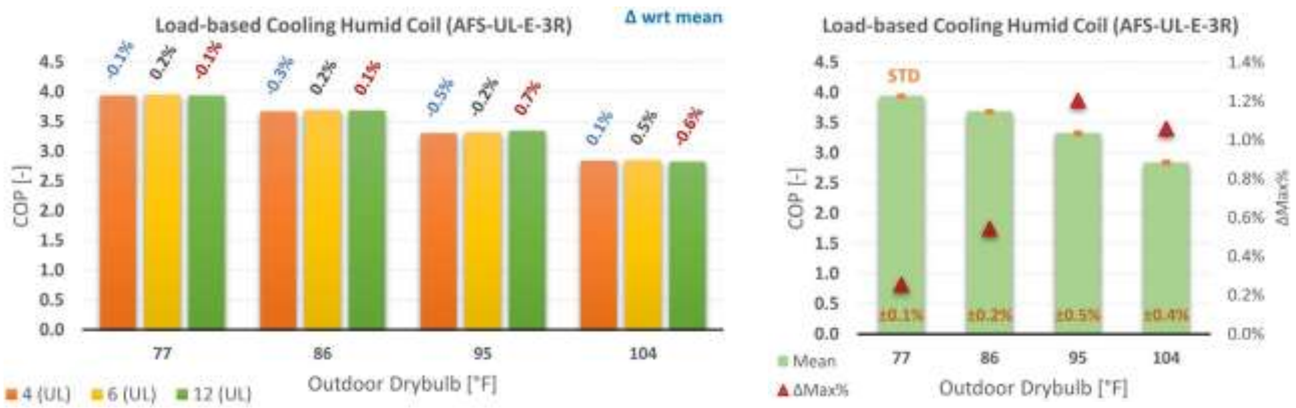


Figure 8. AFS (UL) cooling humid coil test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Figure 7 and Figure 8 show the repeatability comparison of cooling humid coil test intervals performance and unit behavior. Overall good repeatability was observed for all test intervals with a maximum difference in COP between different tests varying from 0.3% to 1.2% and STD from ±0.1% to ±0.5%. Humid coil tests had better repeatability than cooling dry coil tests, which is somewhat counterintuitive because adding humidity to humid coil tests is thought to increase variability due to dynamics associated with the virtual building latent load model, humidity measurement, and test facility controls. The humidity-related variability may have counteracted the ones affecting dry coil tests, resulting in higher repeatability in humid-coil results.

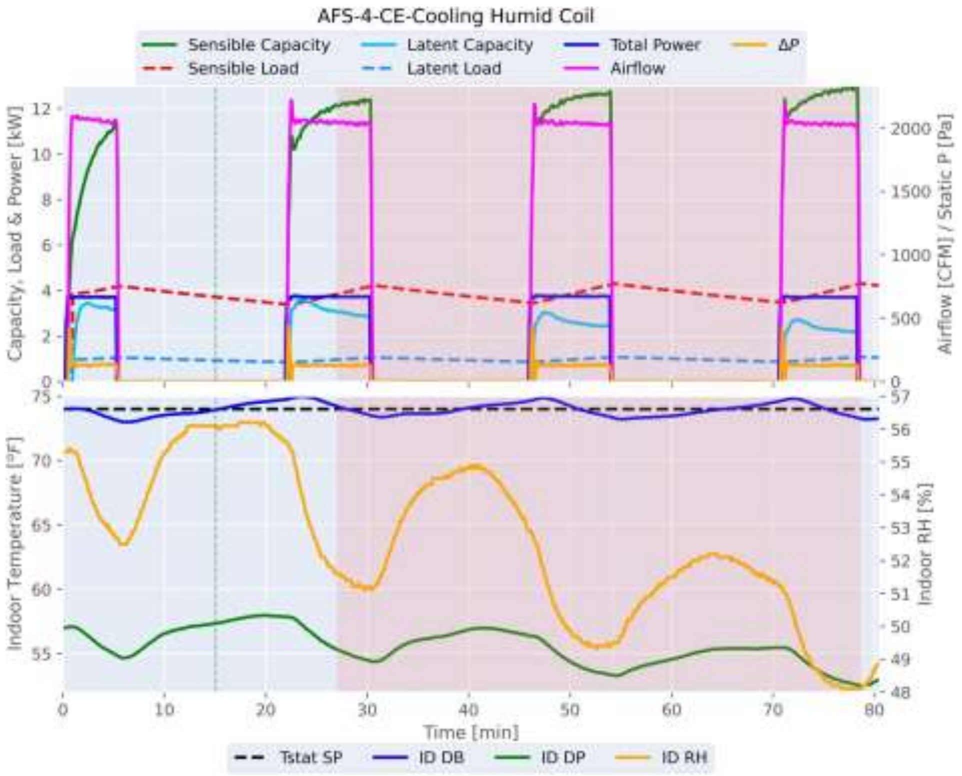


Figure 9. AFS-4 (UL) performance and temperature variations for cooling humid coil test interval CE (77°F)

Another interesting thing was observed in humid coil test intervals where the unit cycled on/off as shown with one example results in Figure 9 for 77°F ambient condition test interval (CE) in test #4. COP converged during the last two on/off cycles, and so did the indoor temperature; however, indoor humidity did not converge and was still decreasing with cycles as can be seen with the indoor relative humidity (ID RH) and indoor dewpoint (ID DP) variation in the bottom subplot. The effect of this can be seen in the decreasing latent cooling rate and increasing sensible cooling rate with cycles. Currently, EXP07 convergence criteria do not consider indoor humidity for convergence, so maybe a discussion is required on whether it should be included or at least some bounds on the indoor humidity should be defined for the UUT to maintain conditions suitable for thermal comfort.

For each test, the cooling seasonal coefficient of performance ($SCOP_C$) was estimated for different climate zones based on the measured performance at cooling dry coil and humid coil test conditions and a temperature bin method as per EXP07:19. Figure 10 shows comparisons of the estimated cooling SCOP for different climate zones based on three different sets of tests for the AFS unit. It is worth noting here that after tests #4 and #6, the test unit was taken out from the test rooms, stored for some time, and then re-installed again before performing test #12. Figure 11 shows the average value of cooling SCOP for repeated tests with 95% CI and STD for different climate zones. In addition, on the right vertical axis, differences in the maximum and minimum values of cooling SCOP are shown as a percentage of its mean value for three tests. For 4 out of 7 climate zones, which utilize humid coil test results in SCOP estimation (Table 6), the maximum difference in cooling SCOP among three tests is within 0.3% and for the other 3 climate zones which utilize dry coil test results, the maximum difference is around 3%. This is due to the larger variations observed in COP for cooling dry coil test results as shown above, especially during the 113°F outdoor condition test interval. Overall, the AFS unit showed good repeatability in cooling load-based test results.

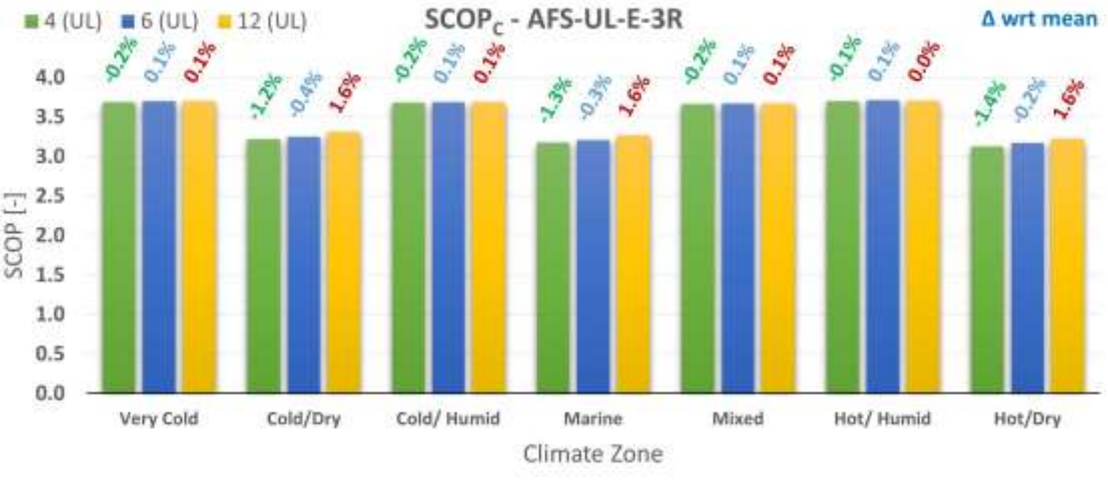


Figure 10. AFS (UL) cooling SCOP comparison in different climate zones for different tests

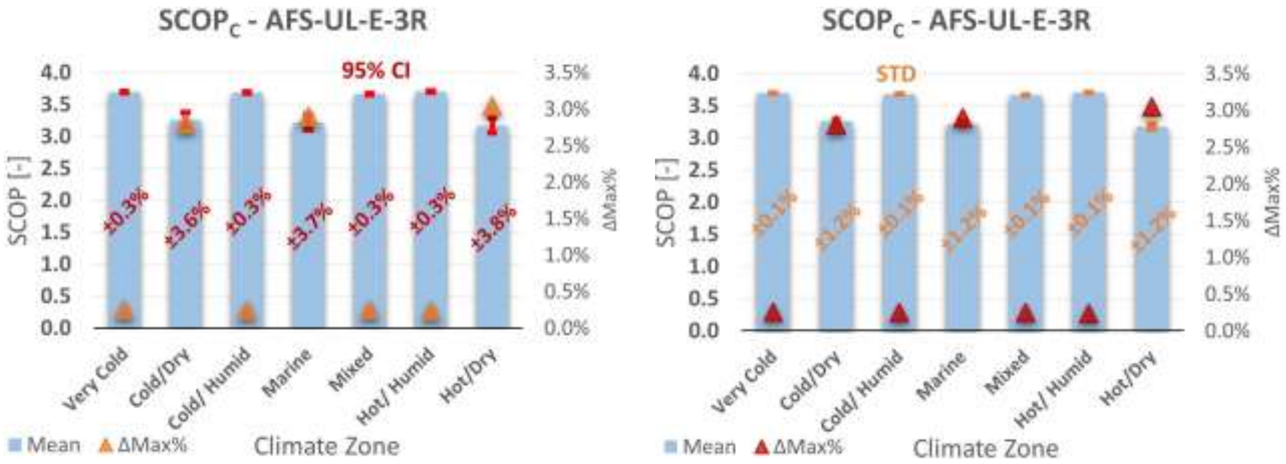


Figure 11. AFS (UL) cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

4.1.2 Heating Mode

Figure 12 and Figure 13 show the repeatability comparison of heating continental test intervals measured COP and unit behavior along with any possible issue (Table 11) with the test interval. The unit cycled on/off at higher ambient conditions with lower building load and then ran out of capacity at ambient temperature 17°F and below where full-load tests were performed. Overall, UUT showed good repeatability between repeated tests in COP as well unit operation mode with a maximum difference in COP varying from 1.1% to 2.5% and STD from ±0.5% to ±1% across 6 test intervals.

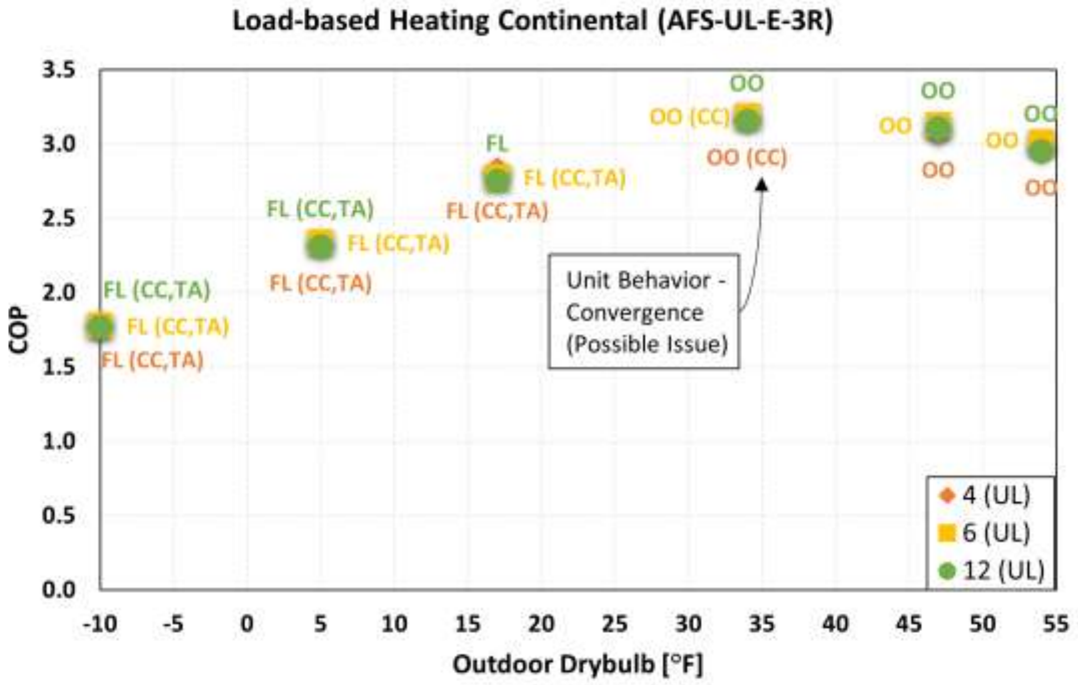


Figure 12. AFS (UL) heating continental test intervals COP for different tests with unit behavior during convergence and possible issues

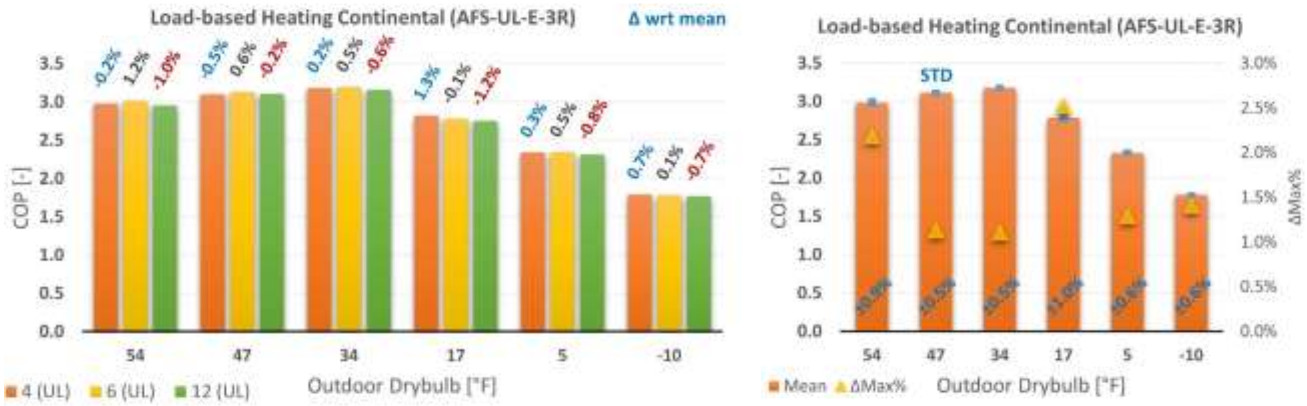


Figure 13. AFS (UL) heating continental test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

However, possible issues were observed related to convergence criteria (CC) and testing approach (TA) in some test intervals, mainly regarding the consideration of defrost during load-based and full-load tests. For example, Figure 14 shows the test unit performance for the heating continental test at 34°F ambient temperature, where UUT showed both defrost as well as on/off cycling behavior together. However, convergence was found with two successive on/off cycles as shown in the highlighted area. Thus, the measured performance for this interval does not account for defrost losses which is not representative of the test unit's overall performance at this test condition. Convergence criteria should be updated to account for defrosts in cases similar to this.

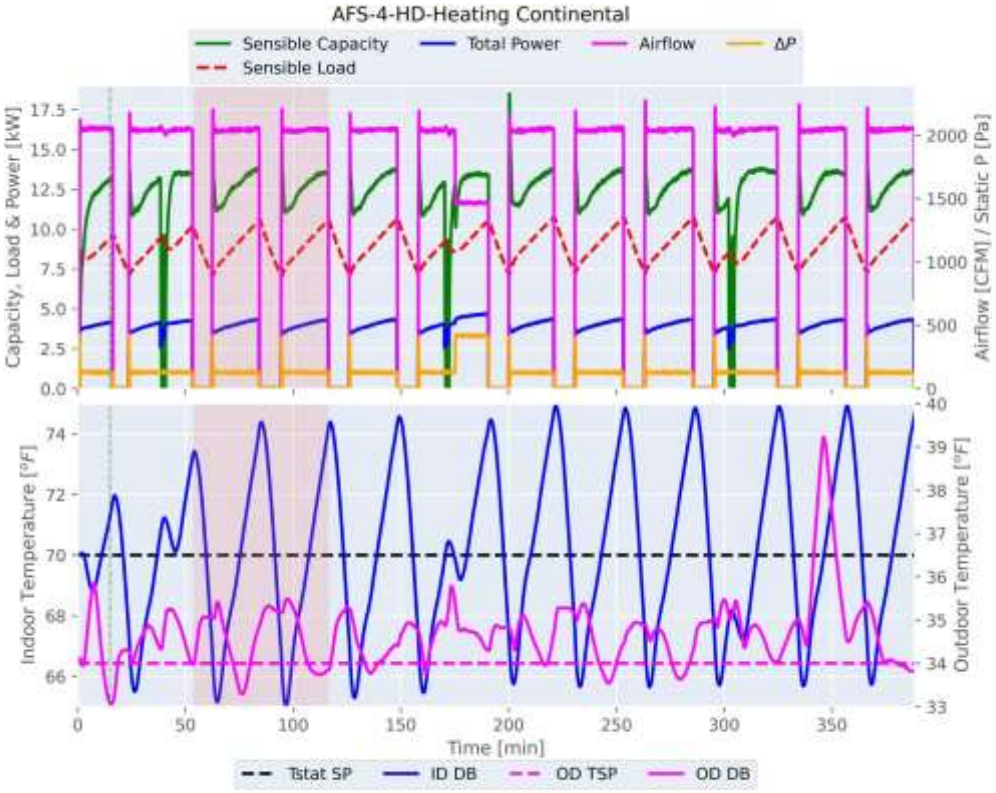


Figure 14. AFS-4 (UL) performance and temperature variations for heating continental test interval HD (34°F)

Further, at the same test condition of 34°F ambient temperature, a similar behavior was observed with on/off cycles and defrosts in test #6; however, in test #12, before the unit went into defrost, convergence was found with on/off cycles as shown in Figure 15. This might lead to the conclusion that there is no defrost at this condition which may not be true as the unit might utilize defrost if it was allowed to run longer. This is a challenge as it will lead to higher performance without accounting for defrosts. A possible solution for this is to update the testing approach by prescribing a minimum but long enough test duration to decide whether the unit utilizes defrost even if convergence is found as in this case. This could be done only for test conditions where defrost from a test unit is expected such as at outdoor temperatures 34°F and below. This will help for cases similar to this; however, it will also increase the test time for cases where the unit doesn't utilize defrost for a test condition.

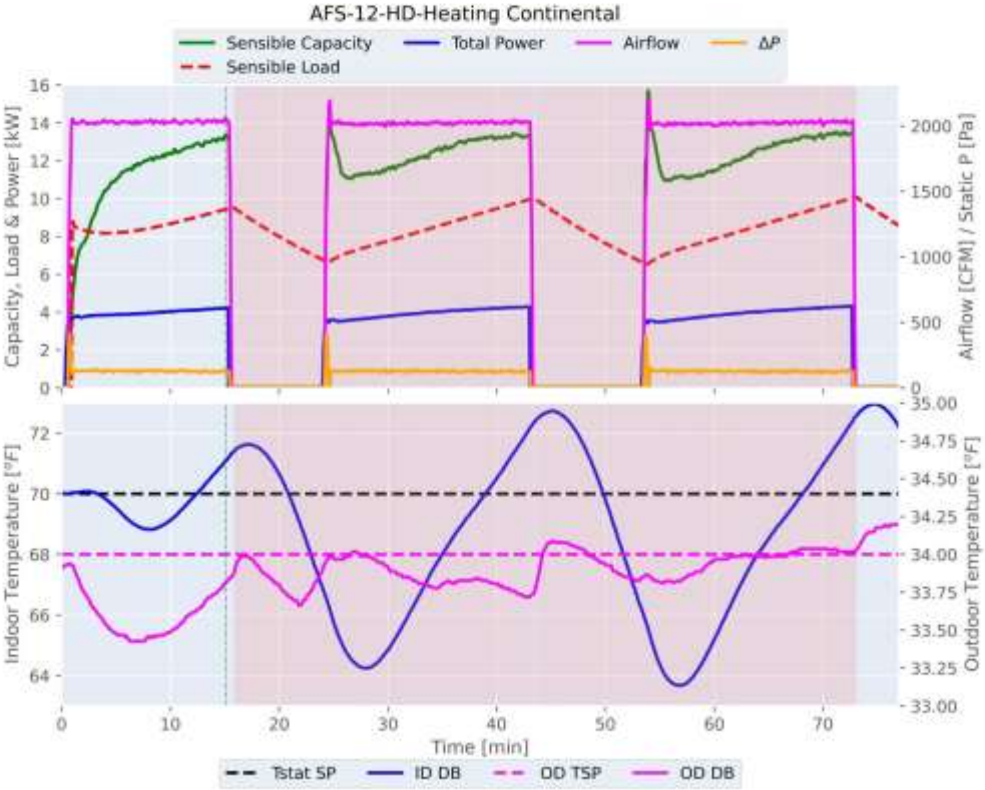


Figure 15. AFS-12 (UL) performance and temperature variations for heating continental test interval HD (34°F)

Somewhat similar convergence criteria and testing approach issue was observed for test intervals at 17°F and below ambient conditions, as shown with an example of test unit performance at 17°F outdoor temperature in Figure 16 for test #4. In this full-load test interval, the unit utilizes defrost around 20 min into the test; however, convergence was found in a steady operation mode around 42-82 min, thus, not accounting for defrost in overall performance. So, similar to as mentioned above, convergence criteria need to be updated to account for defrost as well as the testing approach so that the test is run long enough to measure multiple complete defrost cycles in order to capture overall representative test unit performance for a test condition.

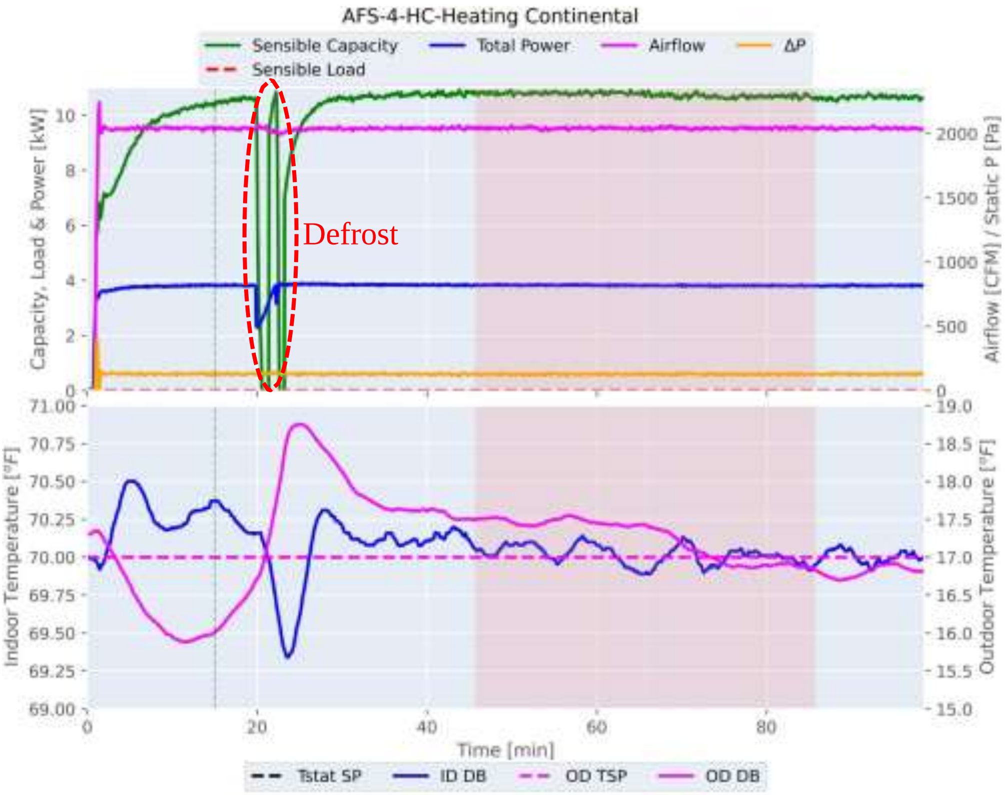


Figure 16. AFS-4 (UL) performance and temperature variations for heating continental test interval HC (17°F)

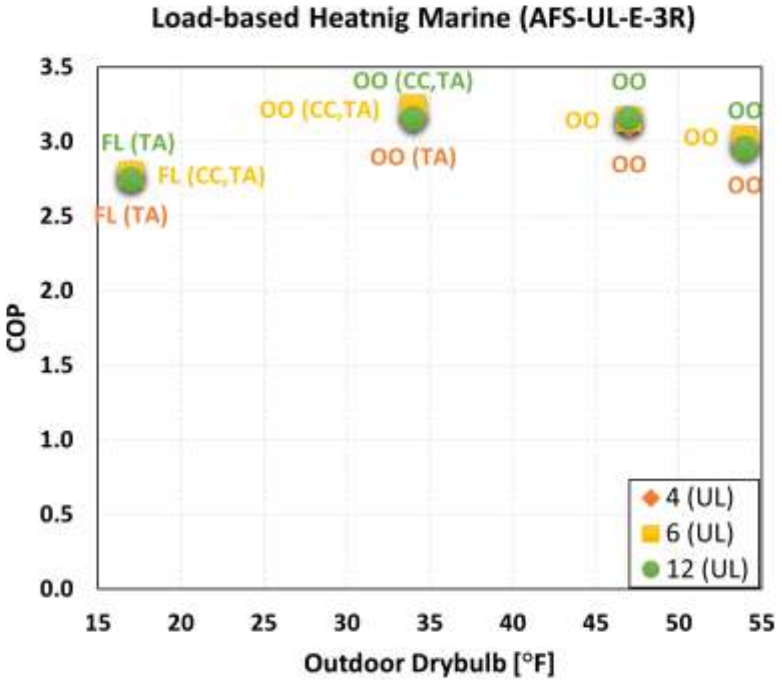


Figure 17. AFS (UL) heating marine test intervals COP for different tests with unit behavior during convergence and possible issues

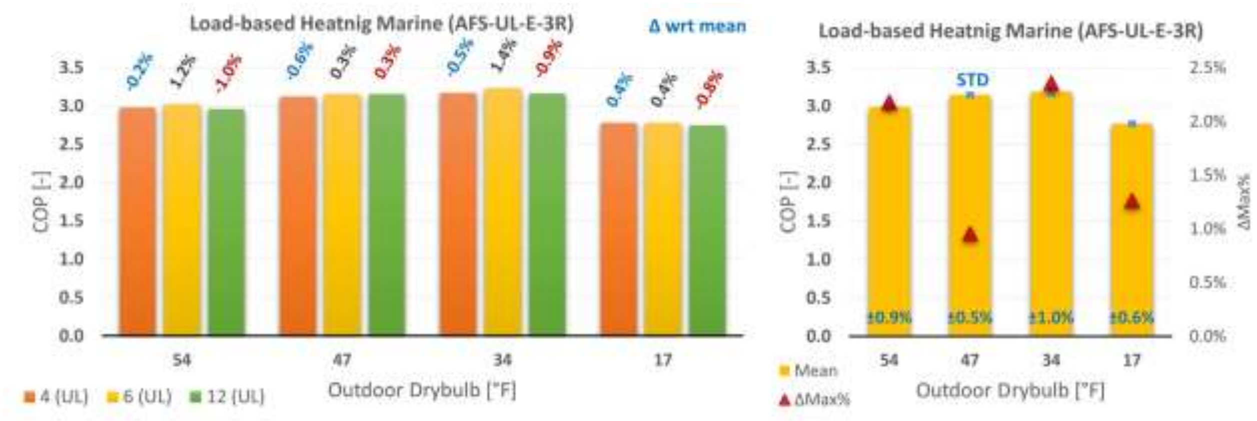


Figure 18. AFS (UL) heating marine test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Figure 17 and Figure 18 show the repeatability comparison of performance measured in heating marine test intervals at 4 different ambient conditions for 3 repeated tests. The test unit showed good repeatability in COP as well as unit behavior with maximum COP difference varying from 1% to 2.4% and STD from ±0.5 to ±1%; comparable to heating continental test interval results. Also, similar possible issues with convergence criteria and testing approach to account for defrost in overall measured performance were observed at 34°F and 17°F ambient condition test intervals.

Figure 19 and Figure 20 show comparisons of the AFS unit estimated heating seasonal coefficient of performance (SCOP) for different climate zones along with the maximum difference and average performance with a 95% CI and STD based on three different sets of tests. In heating, differences in the seasonal performance of different tests are relatively similar and small for different climate zones with a maximum difference of around 1.2%, showing good repeatability in the test results.

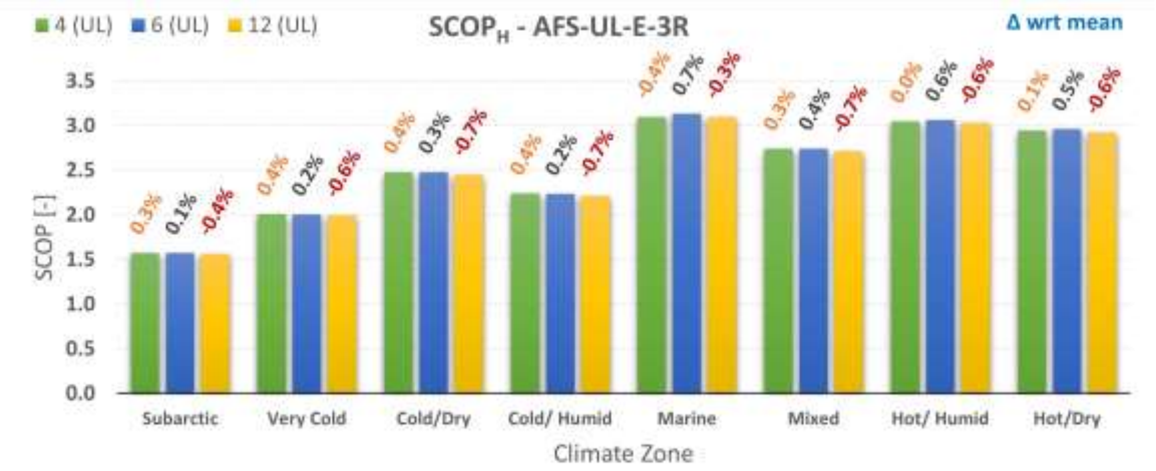


Figure 19. AFS (UL) heating SCOP comparison in different climate zones for different tests

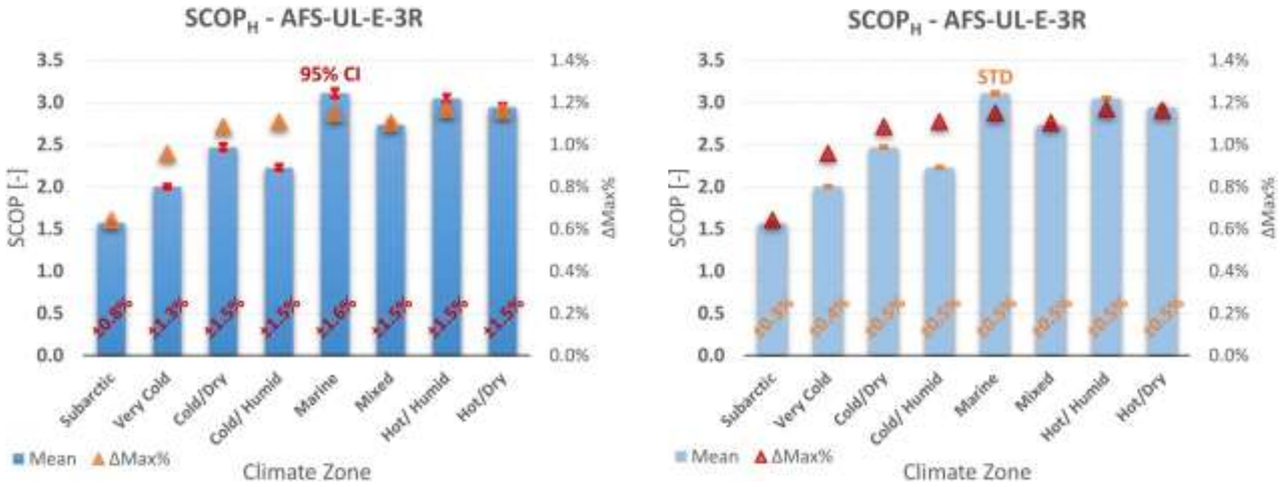


Figure 20. AFS (UL) heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

Table 12 shows the overall SCOP repeatability results summary for AFS with the minimum, maximum, and mean values of three statistical parameters across different climate zones. In summary, the AFS unit shows good repeatability among different test results for cooling as well as heating except for one cooling dry coil test interval (CA), which can be attributed to a failure during one of the tests as described above.

Table 12. AFS-UL-E-3R dataset SCOP repeatability summary results

Metric	No. of Tests	ΔSCOP (Max-Min) [%]			SCOP 95% CI [%]			SCOP STD [%]		
		Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
SCOP$_C$	3	0.2%	3.0%	1.4%	0.3%	3.8%	1.8%	0.1%	1.2%	0.6%
SCOP$_H$	3	0.6%	1.2%	1.0%	0.8%	1.6%	1.4%	0.3%	0.5%	0.5%

4.2 NEEA4 (UL)

In this section, EXP07 repeatability assessment analysis results are presented for the NEEA4 unit, a 1-ton mini-split ductless variable-speed heat pump system, based on three sets of tests from UL for two datasets NEEA4-UL-E-3R and NEEA4-UL-E-2W.

4.2.1 Cooling Mode

Figure 21 and Figure 22 show the COP comparison of cooling dry coil test results with unit operation mode during the convergence period and a possible issue with test interval for three repeated tests in dataset NEEA4-UL-E-3R. Similar unit behavior was observed at the same outdoor temperature test conditions with unit cycling on/off at low ambient temperatures, operating in variable speed steady or hybrid mode at intermediate outdoor temperatures, and running out of capacity at the highest outdoor temperature of 113°F. The maximum difference in COP between three tests under the same test conditions varied from 5.1% to 30.7% and STD from ±2.2% to ±14%, significantly larger compared to the AFS unit. The maximum difference was observed at 104°F outdoor temperature test interval (CB), where a relatively higher COP was measured in test P1 (PREV1) compared to the other two tests, #8 and #10. This was due to the difference in unit dynamic behavior at the same test conditions as shown in Figure 23 for test P1 and in Figure 24 for test #10. At 104°F outdoor temperature, during test P1 unit operated in variable speed steady mode, matching the building load with cooling rate and maintaining indoor temperature near 81°F with relatively low power consumption, whereas, in test #10 it operated in variable speed hybrid mode, ramping up and down with relatively large variation in indoor temperature and power consumption. Convergence was found in both tests as shown with highlighted window. Comparing the converged performance, in test P1, the unit provided around 21.1% lower cooling rate but with around 47.3% lower power consumption rate compared to test #10, resulting in around 30.7% higher COP. In test #10 the convergence was found when the unit was operating at a higher compressor speed which does not seem to be capturing overall representative performance for this test interval. It should include at least one complete ramp up and down cycle, maybe from 35 min to 155 min, or run the test longer until a steady periodic response is achieved, as can be seen in Figure 25 for 95°F outdoor temperature interval (CC) in test P1 with similar convergence criteria issue.

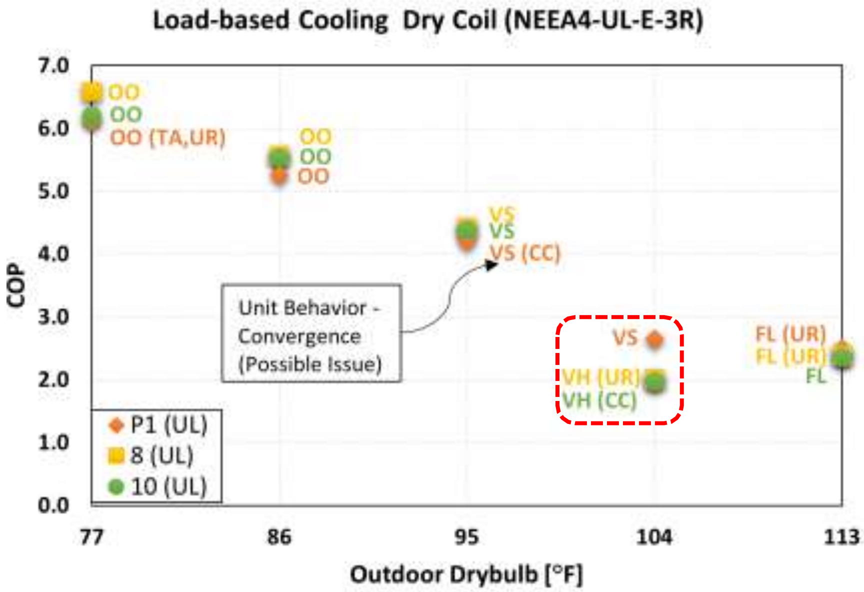


Figure 21. NEEA4 (UL) cooling dry coil test intervals COP for different tests with unit behavior during convergence

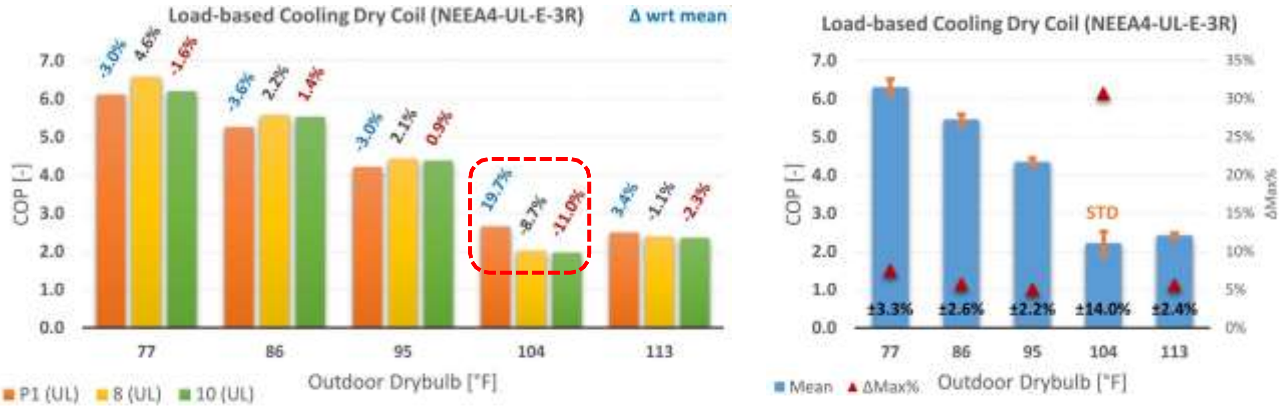


Figure 22. NEEA4 (UL) cooling dry coil test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

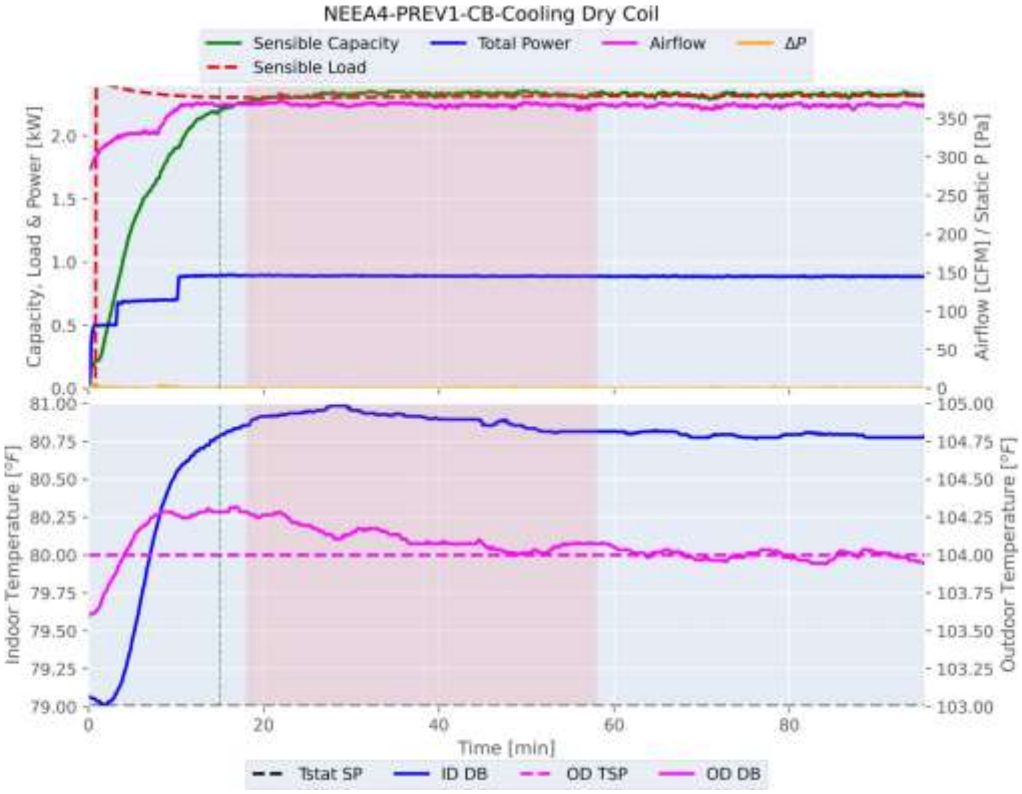


Figure 23. NEEA4-P1 (UL) performance and temperature variations for cooling dry coil interval CB

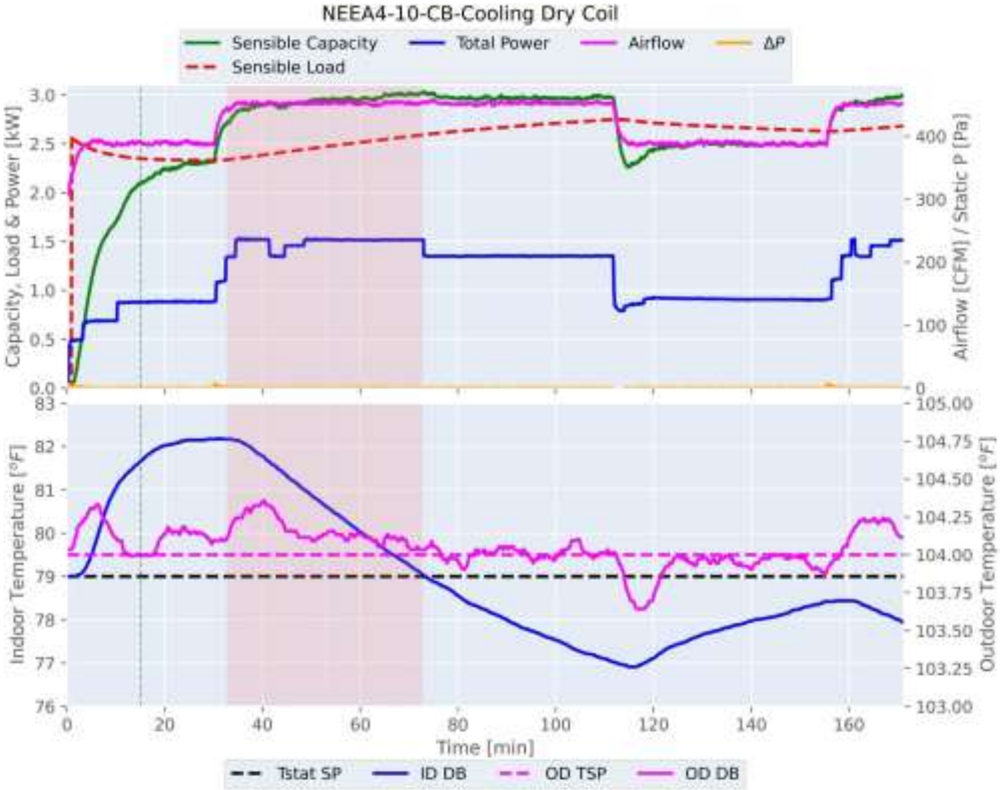


Figure 24. NEEA4-10 (UL) performance and temperature variations for cooling dry coil interval CB

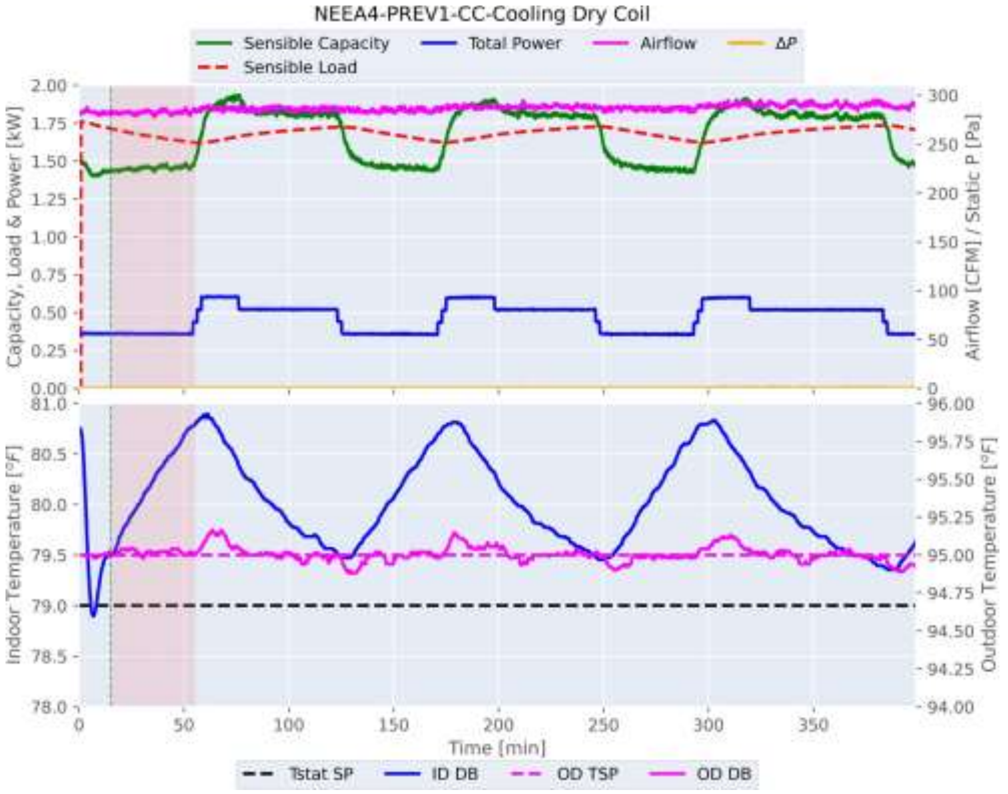


Figure 25. NEEA4-P1 (UL) performance and temperature variations for cooling dry coil interval CC

Figure 26 and Figure 27 show the comparison of test unit performance in cooling humid coil test intervals for three sets of repeated tests at UL. At the same test conditions, a similar unit operation mode was observed during convergence for different tests. The maximum difference in COP between different tests across 4 test intervals varied from 1.1% to 24% and STD ±0.5% to ±11.1%. The major differences were observed for 86°F and 95°F outdoor conditions tests intervals, 22% and 24% respectively, mainly due to higher COP measured in test P1 compared to test #8 and test #10.

Figure 28 and Figure 29 show the test unit performance for the cooling humid coil test interval at 86°F ambient temperature (CD) for tests P1 and #8, respectively. The test unit compressor loads quite differently in the converged window among the two test intervals, which resulted in differences in the cooling rate, power consumption, and indoor temperature. It seems that the differences are mainly due to the variations in the UUT controller response based on its sensing of the space temperature and control design. These differences could also be due to a different thermostat offset, as in test P1, the UUT maintained the room temperature around 75°F with a thermostat setpoint of 74°F without further loading the compressor. Figure 30 and Figure 31 show the test unit performance for the cooling humid coil test interval at 95°F ambient (CC) for tests PREV1 and #8, respectively. Similar to test interval CD (86°F), the test unit behavior among these two test intervals is quite different resulting in a difference in test interval performance and converged indoor average temperature.

It should be noted that test P1 was performed in the first quarter of 2019 at UL, after which the test unit was shipped to PG&E for testing. Then the unit was shipped back to UL and stored for a while before performing tests #8 and #10 in the 2nd and 3rd quarters of 2020. This large time difference and test unit re-installation between these sets of tests could explain the higher performance measured in test P1 compared to the other two, as well as better repeatability among tests #8 and #10, which could be due to variability in the test setup, charge, instrumentation, and testing personnel. However, this large difference is observed only at test intervals where the unit operated in variable-speed mode, not at 77°F ambient condition test interval where the unit cycled on/off, and neither at 113°F outdoor temperature condition with full-load tests. This suggests that this large variation at intermediate ambient temperature conditions could be mainly due to differences in test unit dynamic response (compressor loading/unloading) which mainly depends on test unit control and thermostat sensing rather than other variables as mentioned above.

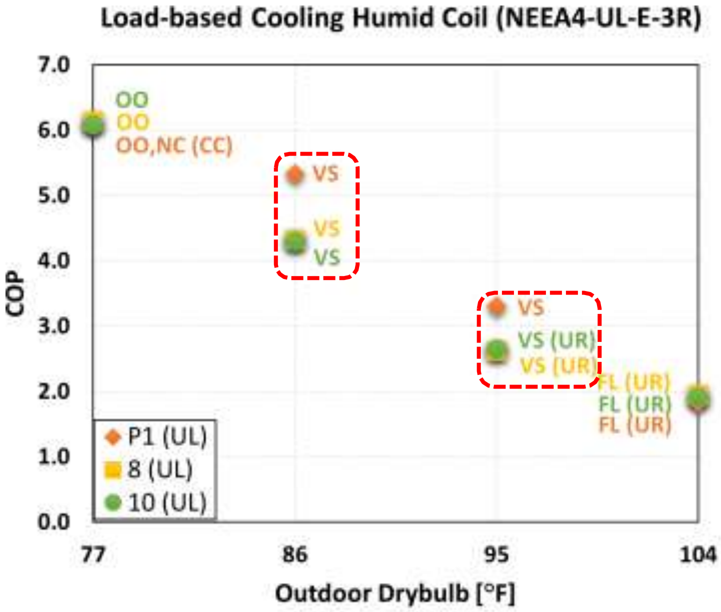


Figure 26. NEEA4 (UL) cooling humid coil test intervals COP for different tests with unit behavior during convergence

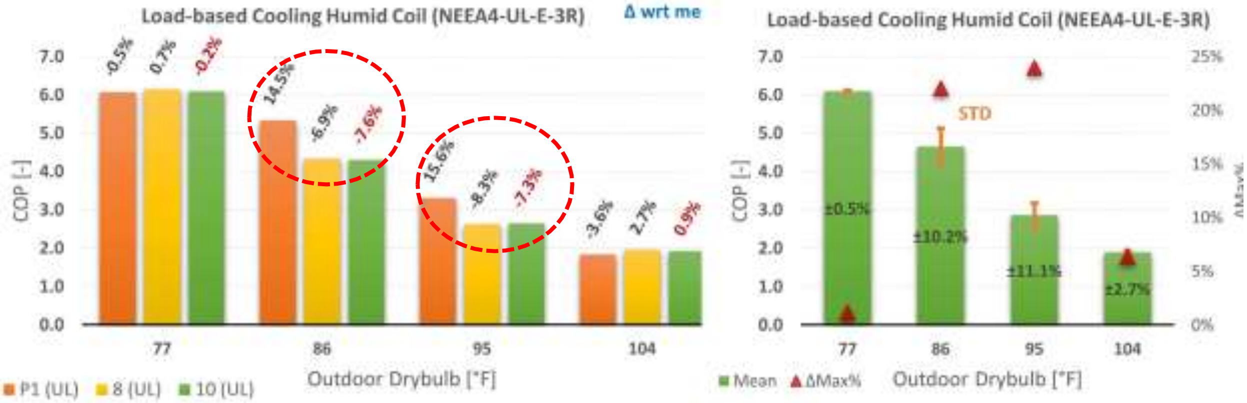


Figure 27. NEEA4 (UL) cooling humid coil test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

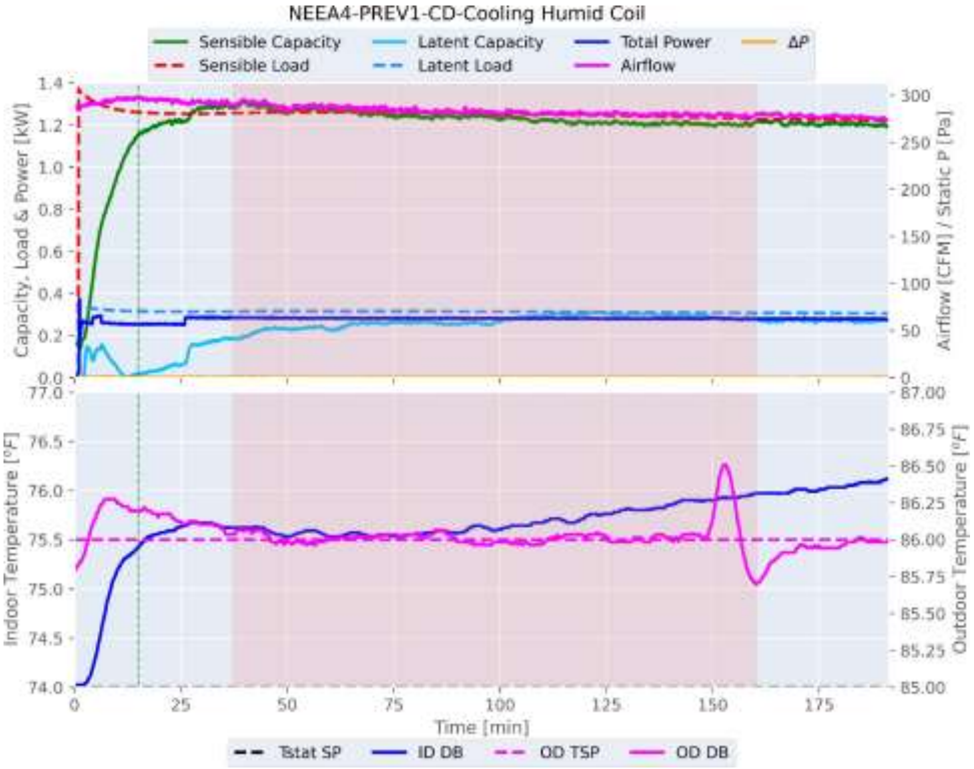


Figure 28. NEEA4-P1 (UL) performance and temperature variations for cooling humid coil test interval CD (86℉)

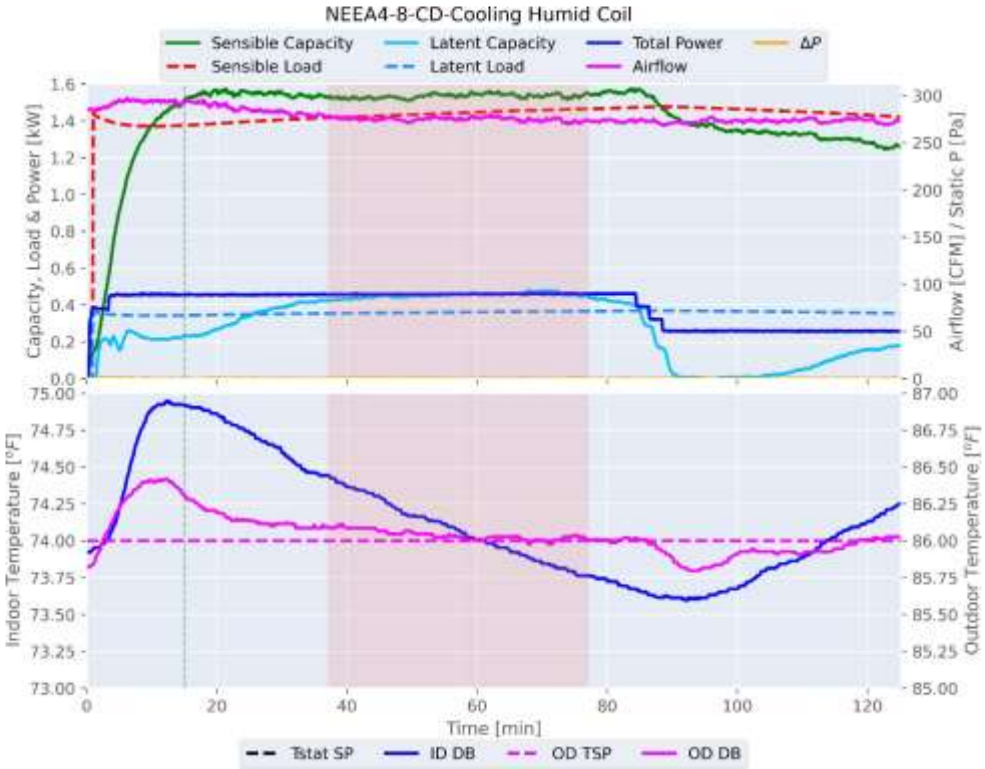


Figure 29. NEEA4-8 (UL) performance and temperature variations for cooling humid coil test interval CD (86℉)

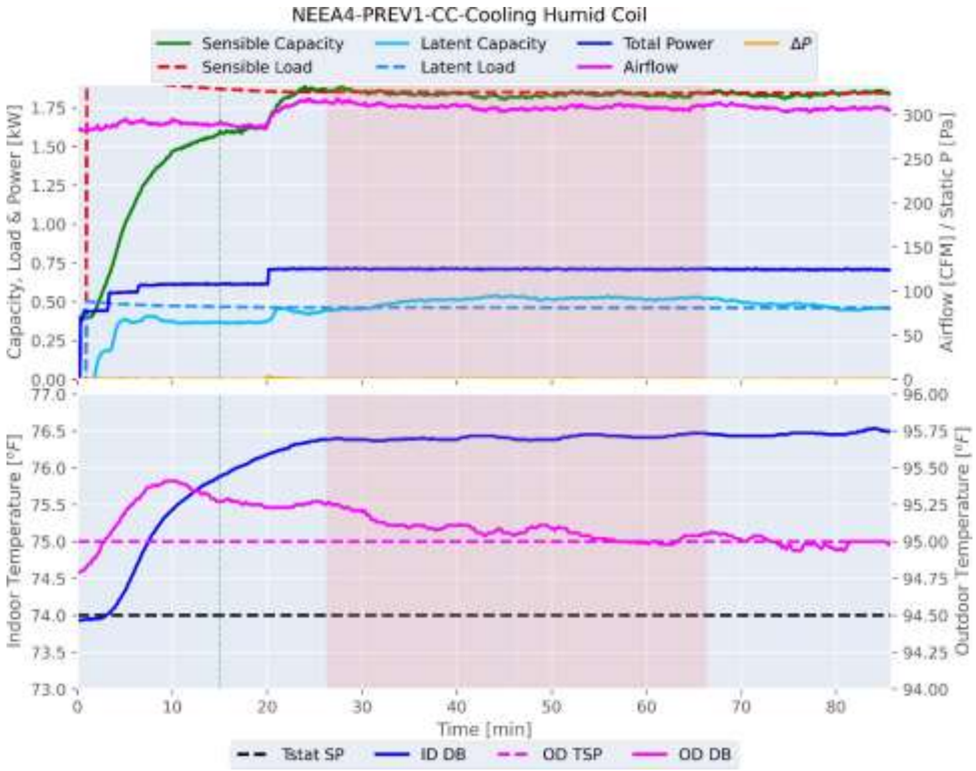


Figure 30. NEEA4-P1 (UL) performance and temperature variations for cooling humid coil test interval CC (95℉)

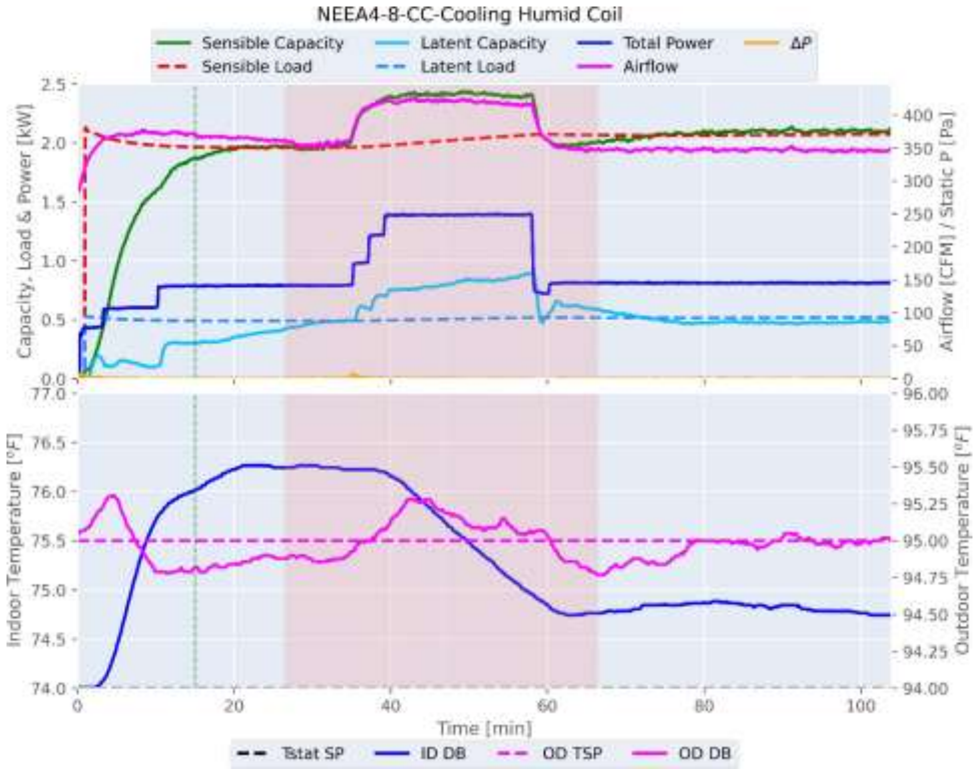


Figure 31. NEEA4-8 (UL) performance and temperature variations for cooling humid coil test interval CC (95℉)

Figure 32 and Figure 33 show the comparison of NEEA4 estimated cooling SCOP for different climate zones based on measured performance in three sets of tests at UL. For climate zones that utilize cooling dry coil test results in SCOP estimation (3 out of 7), overall good repeatability is observed in cooling SCOP with maximum difference varying from 1.5% to 4.5% and STD from ±0.7% to ±1.9%. However, for climate zones utilizing humid coil test results, a larger variation in cooling SCOP was observed with maximum difference varying from 13.5% to 14.3% and STD from ±6.3% to ±6.7%. This was mainly due to higher SCOP estimation based on test P1 results in which relatively higher COP was measured for two test intervals as discussed above.

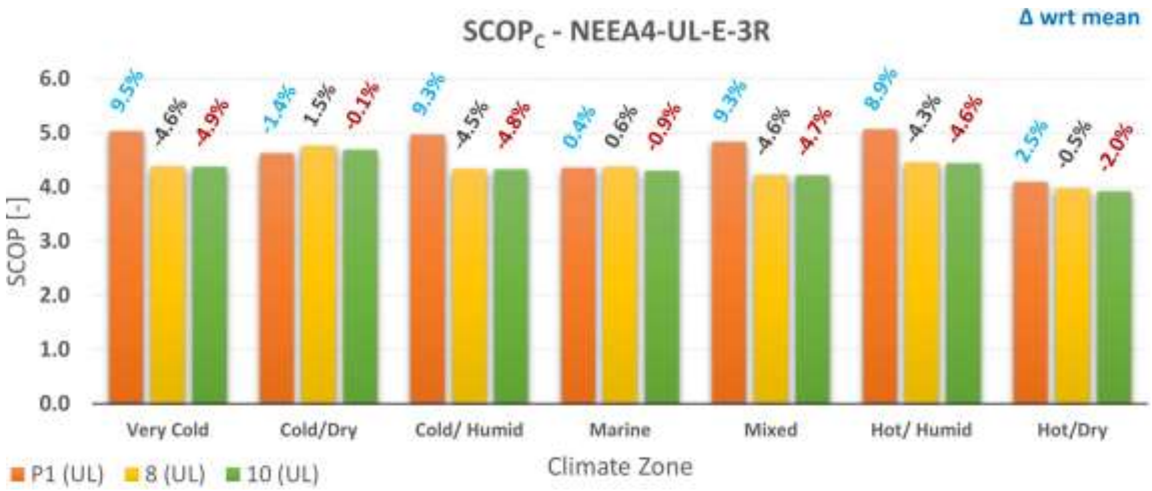


Figure 32. NEEA4 (UL) Cooling SCOP comparison in different climate zones for different tests

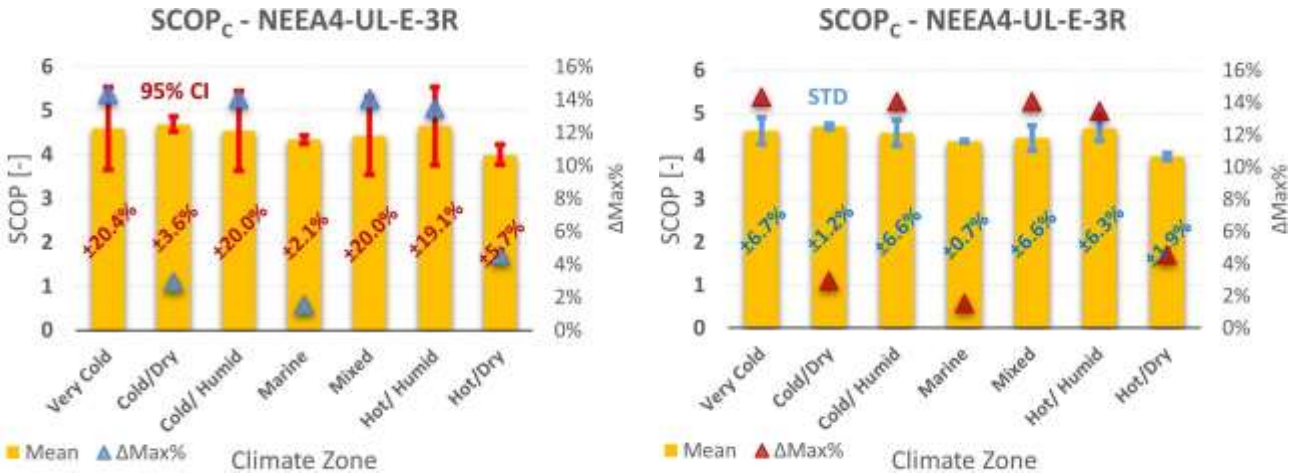


Figure 33. NEEA4 (UL) cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

As shown above, the largest variation in cooling repeatability mainly resulted from test P1 which was done around 16 months before the other two tests, and the unit was moved between labs and re-installed in between. Repeatability improves significantly if only tests #8 and #10 are compared without considering test P1 for dataset NEEA4-UL-E-2W as shown in Figure 34 and the maximum SCOP difference varies only between 0.2% to 1.6% and STD from ±0.1% and ±0.8% across different climate zones.

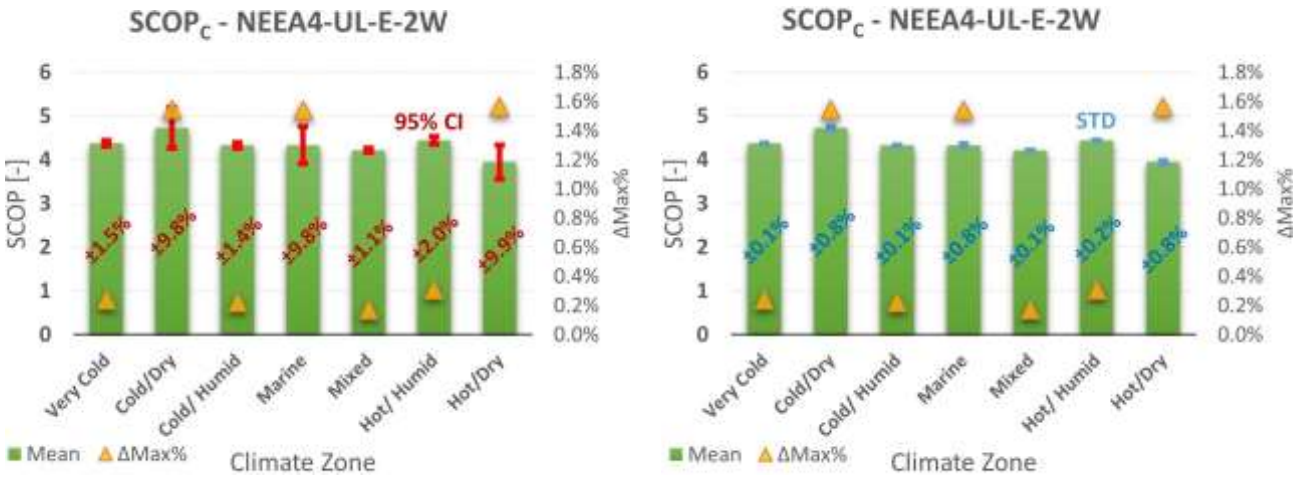


Figure 34. NEEA4 (UL) cooling SCOP average and maximum difference for repeated tests with dataset NEEA4-UL-E-2W with a) 95% confidence interval, and b) standard deviation

4.2.2 Heating Mode

Figure 35 and Figure 36 show the comparison of test unit performance in heating continental test intervals for 3 repeated tests at UL in dataset NEEA4-UL-E-3R. At the same test conditions, even though in general similar unit operation modes were observed, a quite different unit dynamic response was observed in some test intervals among the three tests. One interesting thing to notice is that in none of these heating tests, did convergence criteria include the defrost performance. In almost all the test intervals some possible issues were noted related to either convergence criteria, testing approach, or inconsistent unit dynamic response. The convergence criteria issues mainly related to its not capturing the overall representative performance during; i) unit on/off cycling behavior with a mix of short and long cycles; ii) variable speed hybrid mode where unit ramps up and down without cycling off, and iii) unit utilizing defrost and convergence criteria didn't include that in overall test interval performance. The testing approach issue is related to the case when UUT was going into defrost and the test was not run long enough to capture enough complete defrost cycles to capture overall representative performance at that test condition.

The maximum difference in COP among 3 tests for different test intervals varies from 8.4% to 32.9% and STD from ±3.6% to ±13.9%. Overall, poor repeatability was observed in heating continental test intervals. There were relatively large variations in measured performance for 54°F, 47°F and 17°F outdoor temperature test intervals. Even in the full-load test at -10°F ambient temperature, the maximum difference in COP was around 15.4% between tests P1 and #8.

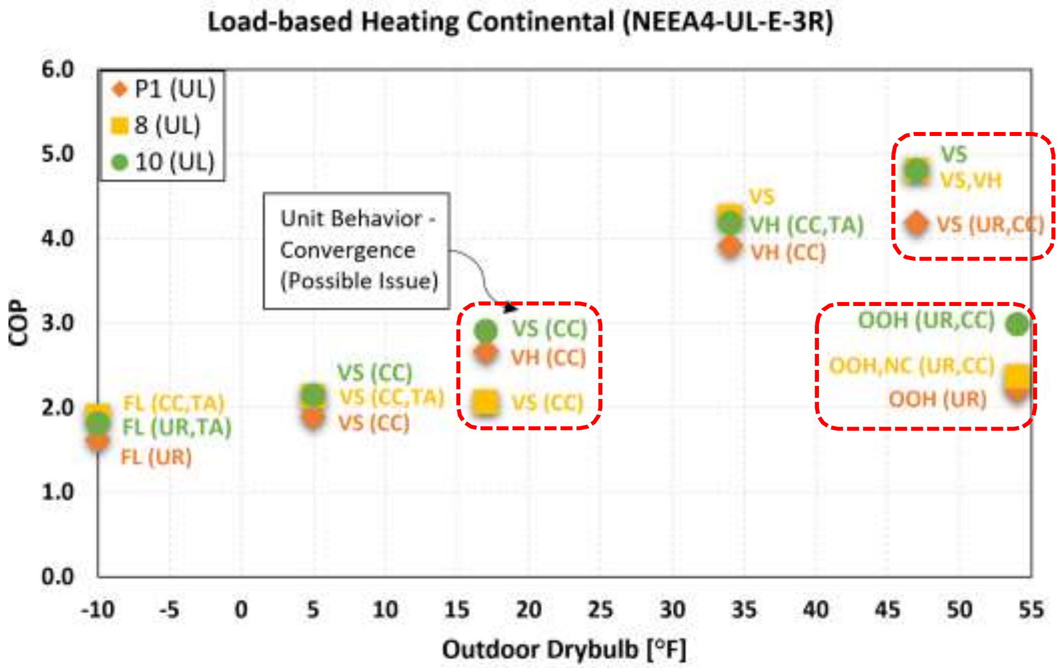


Figure 35. NEEA4 (UL) heating continental test intervals COP for different tests with unit behavior during convergence

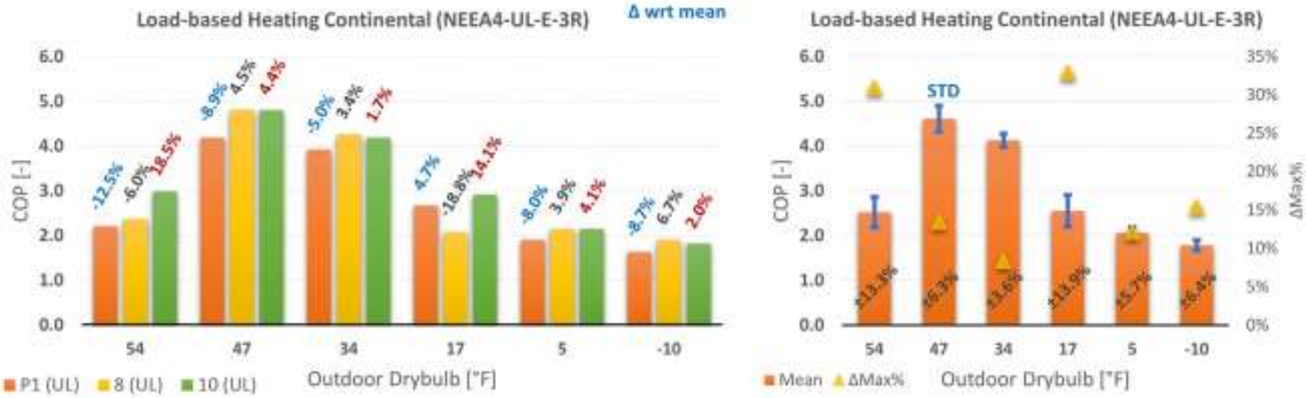


Figure 36. NEEA4 (UL) heating continental test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

For 54°F test interval (HF), the differences in performance for tests P1, #8, and #10 are due to differences in unit cycling behavior, indoor temperature variation, and convergence period as shown in Figure 37, Figure 38, and Figure 39. This could be mainly due to test unit inconsistent dynamic behavior at same tests conditions in different tests. In test P1, the unit only showed short on/off cycling behavior with a small variation in indoor temperature, whereas, in tests #8 and #10, a mix of short and long on/off cycles were observed. Further, in test #10, the convergence was found in shorter on/off cycles in the beginning and the longer on/off cycles were not included in the overall performance, but still resulted in relatively higher COP compared to tests P1 and #8. Even in tests #8 and #10, which were performed relatively close in time without re-installation, a different cycling behavior was observed indicating the unit's inconsistent dynamic behavior under the same test conditions.

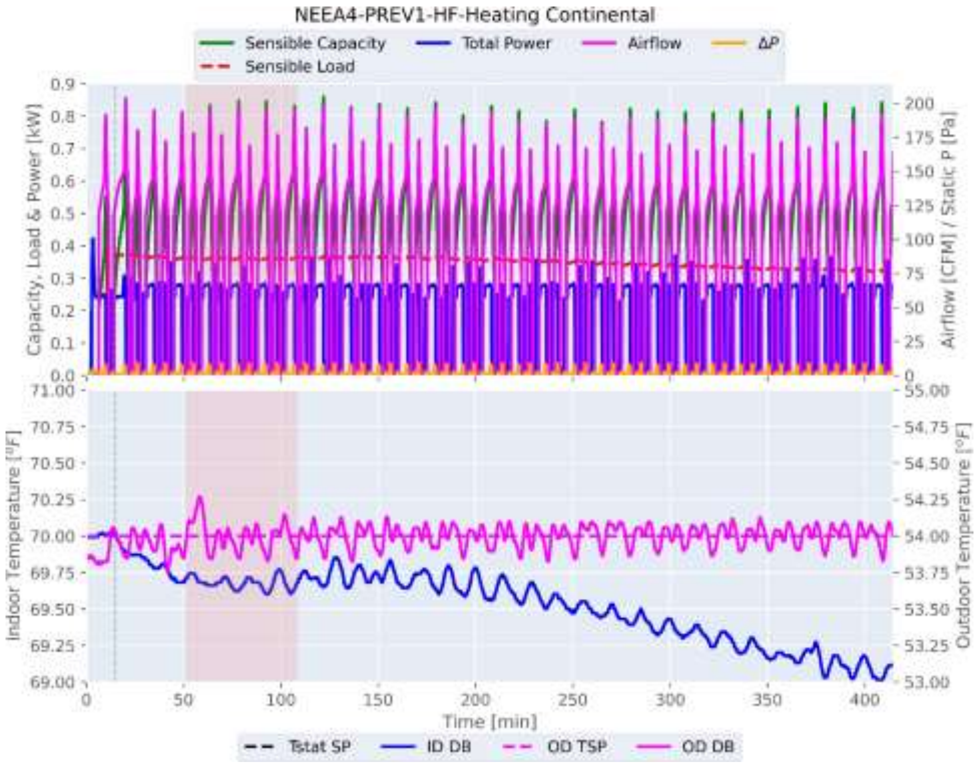


Figure 37. NEEA4-P1 (UL) performance and temperature variations for heating continental test interval - HF (54℉)

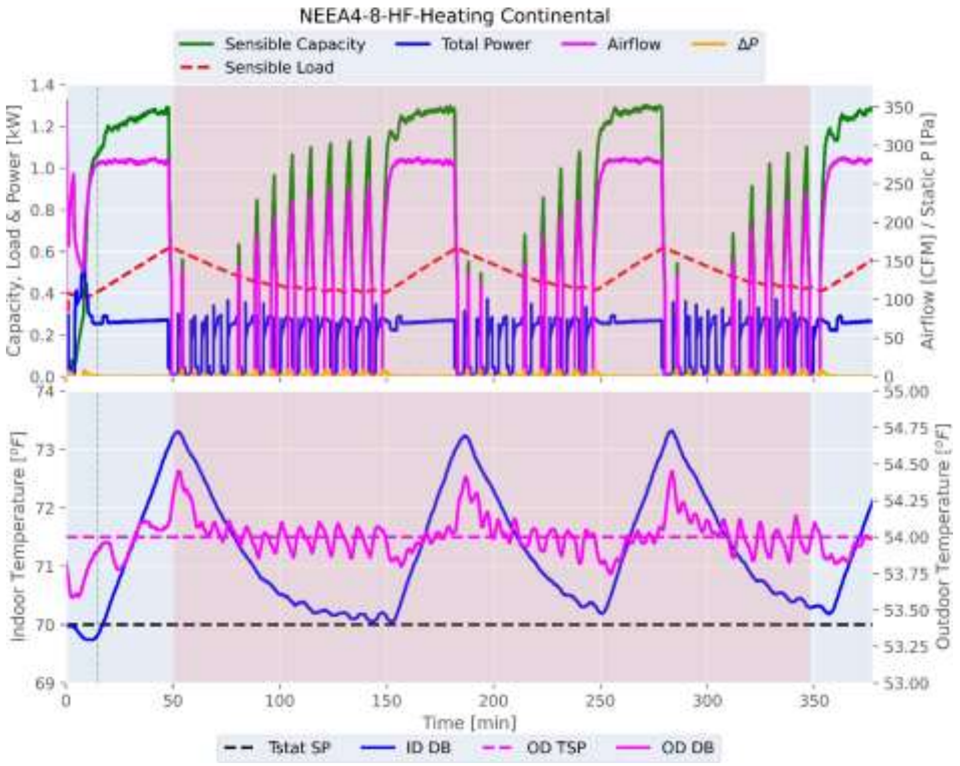


Figure 38. NEEA4-8 (UL) performance and temperature variations for heating continental test interval - HF (54℉)

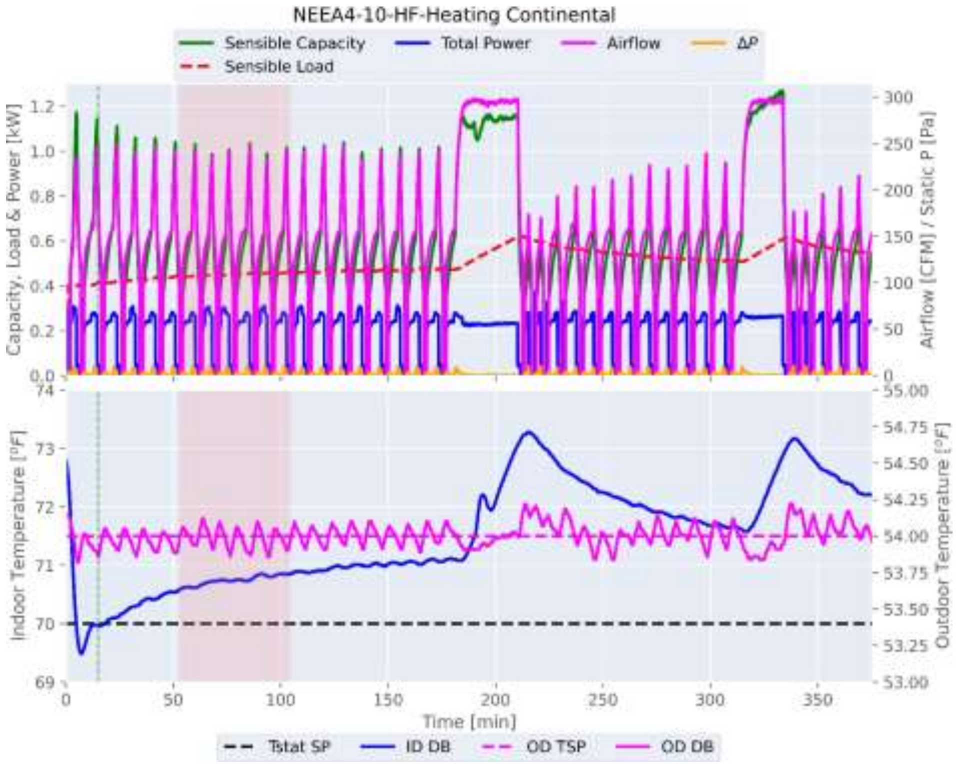


Figure 39. NEEA4-10 (UL) performance and temperature variations for heating continental test interval - HF (54°F)

In the 47°F outdoor temperature test interval, around 13% lower COP was measured in test P1 compared to tests #8 and #10 with similar performance. Figure 40 and Figure 41 show the test unit performance with time at 47°F ambient temperature for tests P1 and #8, respectively. During test P1, the test unit was cycling on/off, but the convergence occurred in steady operation mode before the unit cycled off for the first time. In test #8 variable-speed behavior was observed, possibly due to the shorter test duration. Convergence in both tests captured steady operation mode performance with around 19% lower heating rate in test P1, but with only 6.5% lower power consumption. This results in around 13.4% lower COP in test P1 compared to test #8. This difference could be because of differences in test unit dynamic behavior, test setup including instrumentation and charge, and difference in indoor temperature. Furthermore, the current convergence criteria as per EXP07 should be investigated to make sure that representative test unit performance is captured for cycling on/off cases similar to test P1.

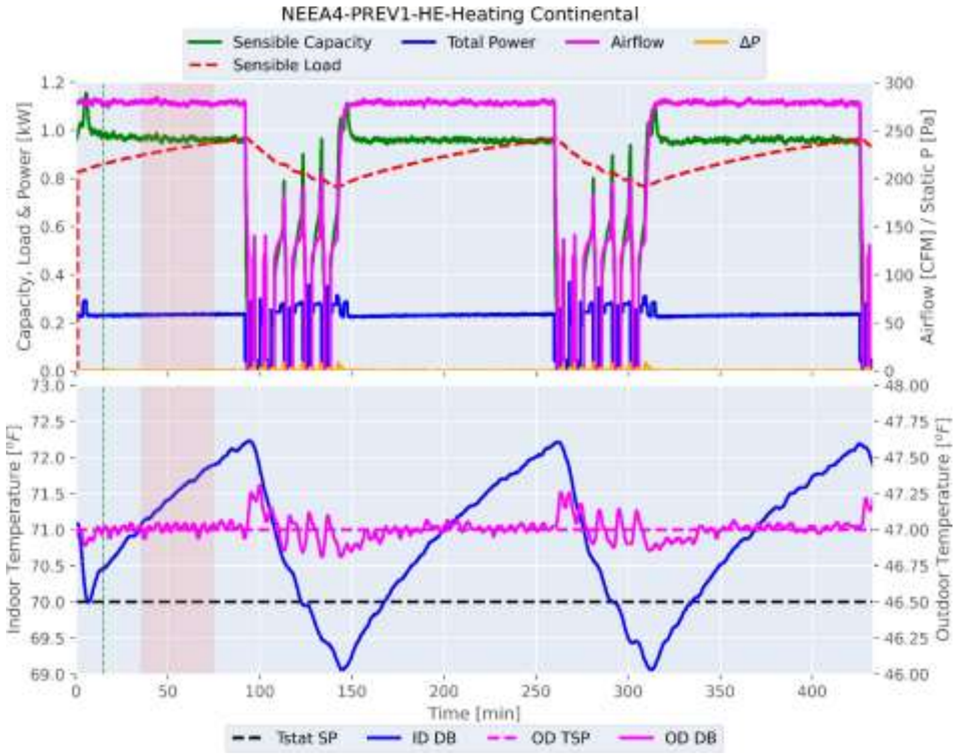


Figure 40. NEEA4-P1 (UL) performance and temperature variations for heating continental test interval HE (47℉)

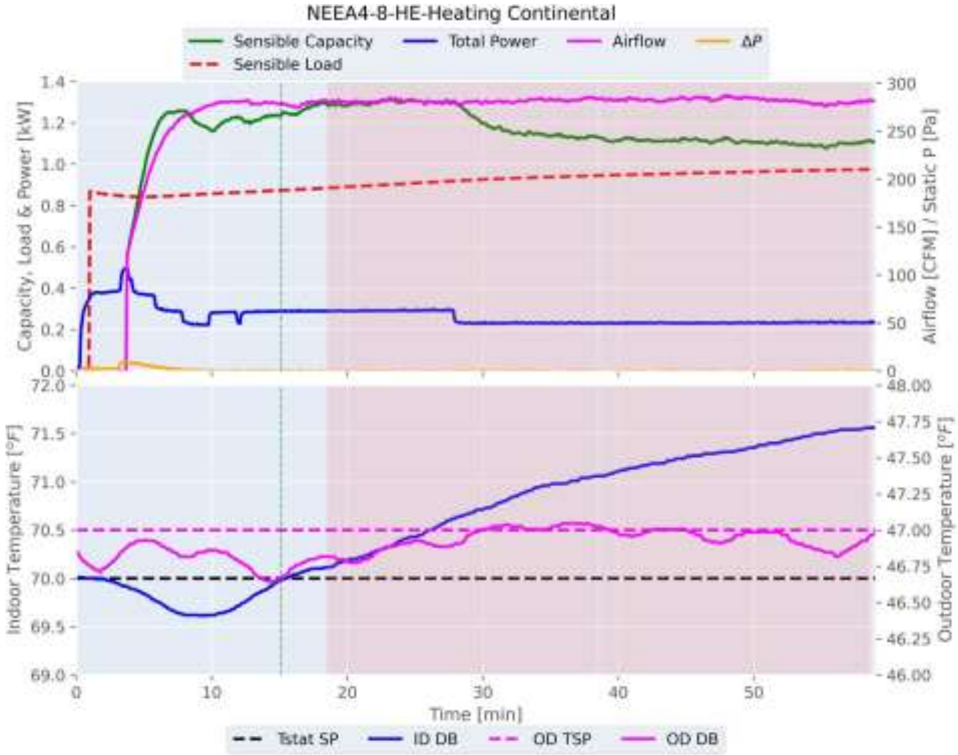


Figure 41. NEEA4-8 (UL) performance and temperature variations for heating continental test interval HE (47℉)

For the 17°F heating continental test interval (HC), the differences in performance are mainly due to the occurrence of convergence at different periods within each test where the unit was modulating its compressor without cycling off as can be seen in Figure 42, Figure 43, and Figure 44. For tests P1 and #10, the convergence occurred near the valley of total power; whereas, for test #8 it occurred close to a peak of the modulating cycle, resulting in comparatively lower performance for test #8. In order to address this case of a variable speed hybrid cycle, it is important to update the current convergence criteria. The other thing to notice is that when the unit was run longer at this test interval in test P1, there was a defrost around 4 hours into the test, which, based on the current convergence criteria in CSA EXP07:19, will be ignored if convergence occurs before that. A further investigation is recommended to decide whether a test unit should be run for a minimum period to see whether a defrost cycle occurs or not and if it occurs then how it should be included in the performance measurement for that test without increasing the test time unreasonably.

Similar behavior of variable speed hybrid cycle was observed in 34°F test interval (HC), where defrost was only observed in test #10 out of three tests at this low ambient temperature condition, which seems odd. In the test interval at 5°F ambient temperature (HB), even though the test unit went into defrost, the convergence was found in the steady operation in between and the defrost was not included in overall performance. Also, the test was not run long enough to capture multiple defrosts as convergence was already found. Similar behavior and issue were observed in the full-load test at -10°F test interval. In these two lowest ambient temperatures test intervals, relatively lower performance was observed in test P1 compared to the other two tests with similar performances.

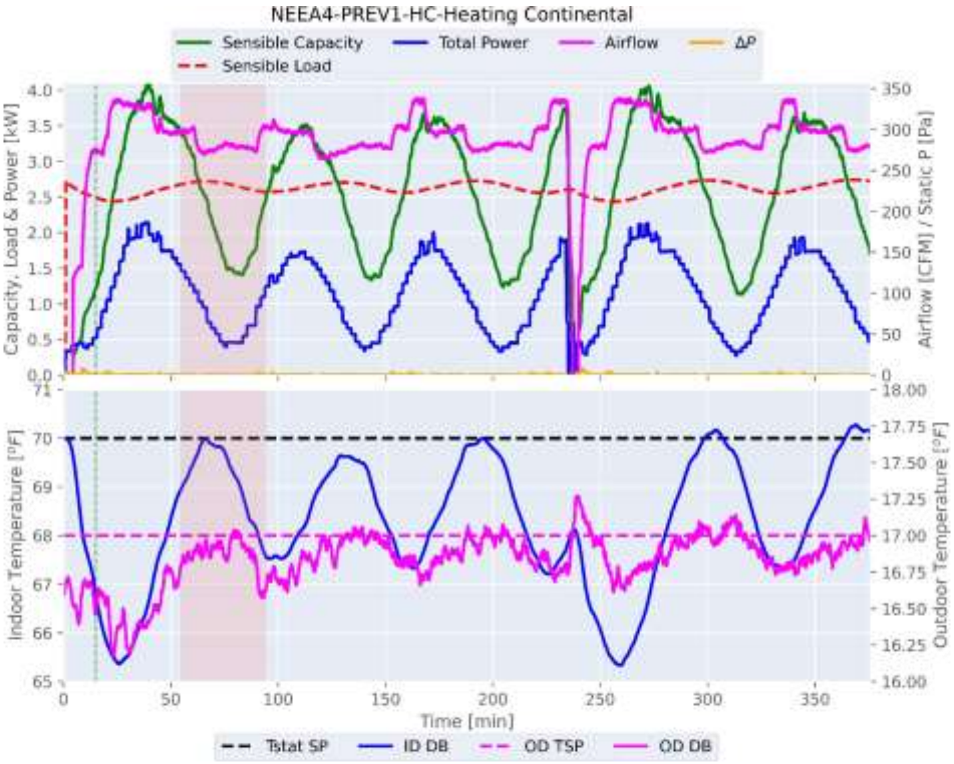


Figure 42. NEEA4-P1 (UL) performance and temperature variations for heating continental test interval HC (17°F)

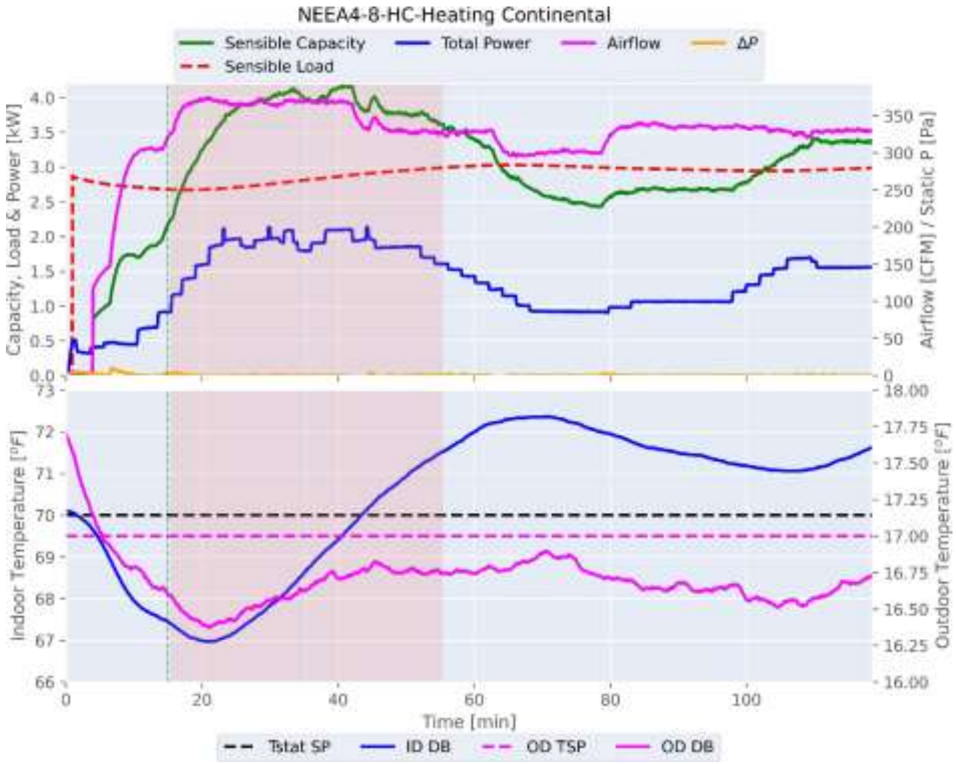


Figure 43. NEEA4-8 (UL) performance and temperature variations for heating continental test interval HC (17℉)

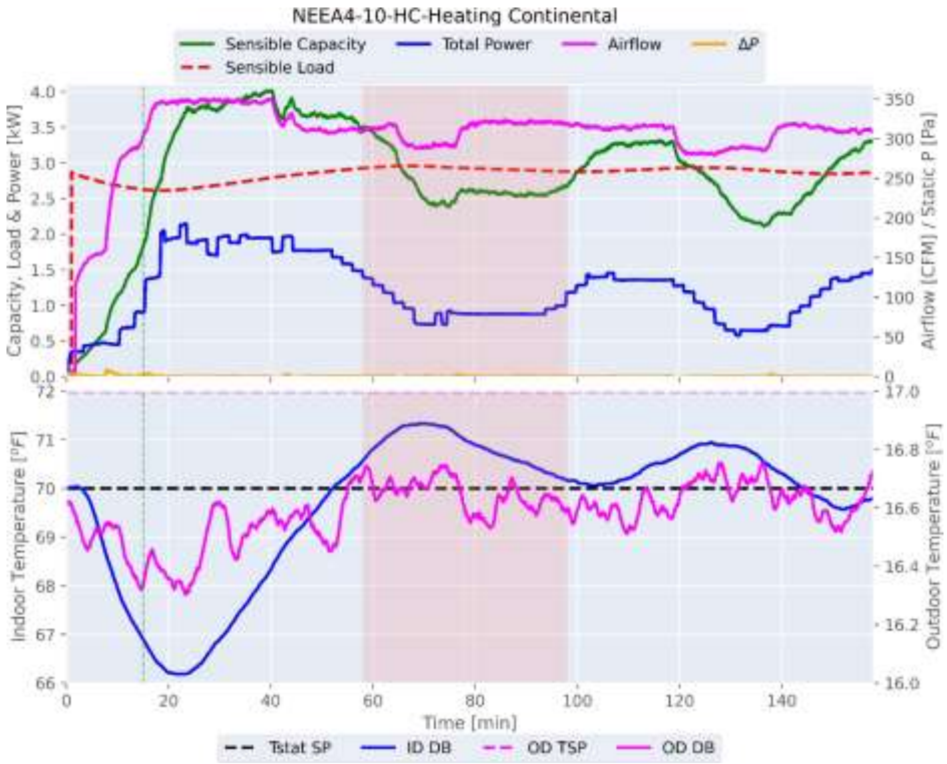


Figure 44. NEEA4-10 (UL) performance and temperature variations for heating continental test interval HC (17℉)

Figure 45 and Figure 46 show the repeatability comparison of test unit performance in heating marine test intervals. Similar to heating continental test conditions, overall poor repeatability was noted, and the lowest COPs were measured in test P1 intervals among three tests. The maximum difference in COP varies from 4.8% to 22.3% and STD ±2% to ±9.1%, marginally better than continental test intervals. In heating marine test intervals, similar possible issues related to the convergence criteria during short cycling, variable-speed hybrid mode, and defrost were observed as described above for continental conditions. Also, defrost performance was not captured in test intervals at 34°F and 17°F outdoor temperatures.

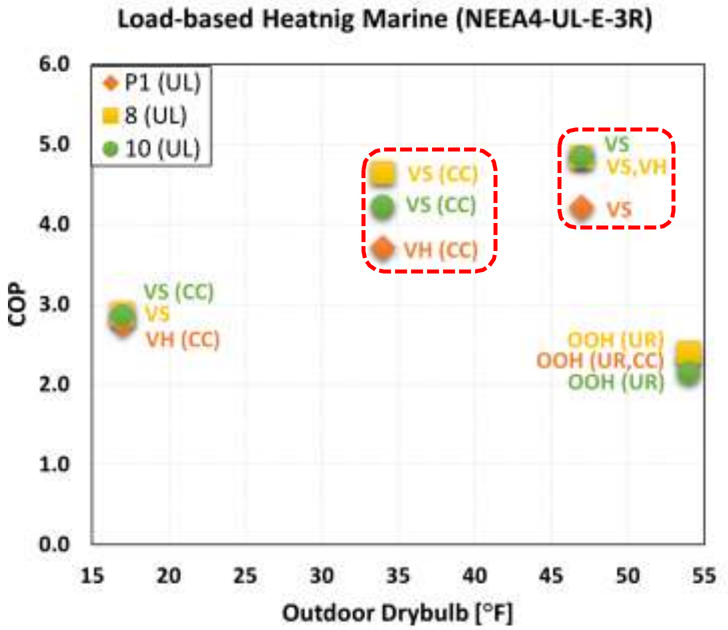


Figure 45. NEEA4 (UL) heating marine test intervals COP for different tests with unit behavior during convergence

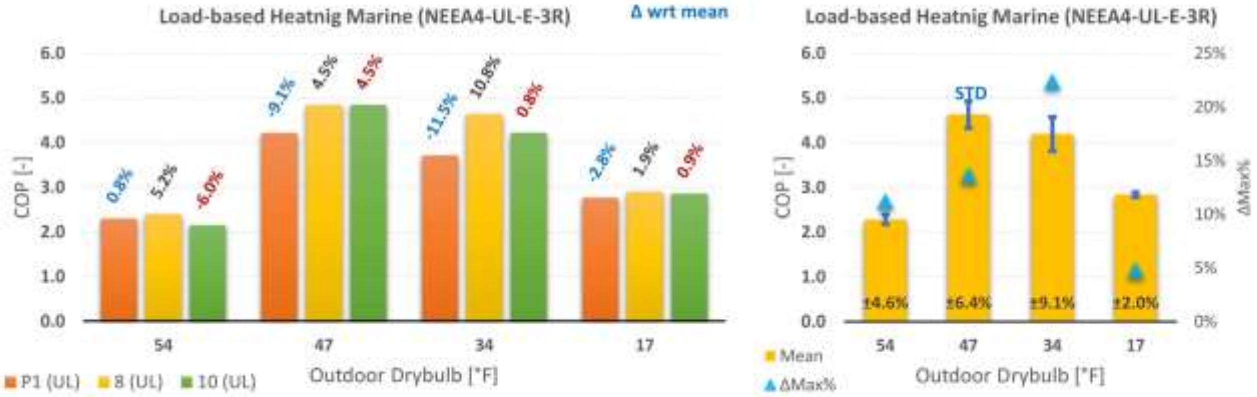


Figure 46. NEEA4 (UL) heating marine test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Figure 47 and Figure 48 show comparisons of NEEA4 unit estimated heating SCOP for different climate zones along with the maximum differences and average performances with a 95% CI and STD based on three sets of tests at UL. The differences in heating SCOP among the three tests are relatively large and consistent among different climate zones, with maximum differences varying from 12.2% to 17.3% and STD from ±5.1% to ±7.1%. In contrast to cooling, unit heating performance in test P1 was relatively low compared to tests #8 and #10.

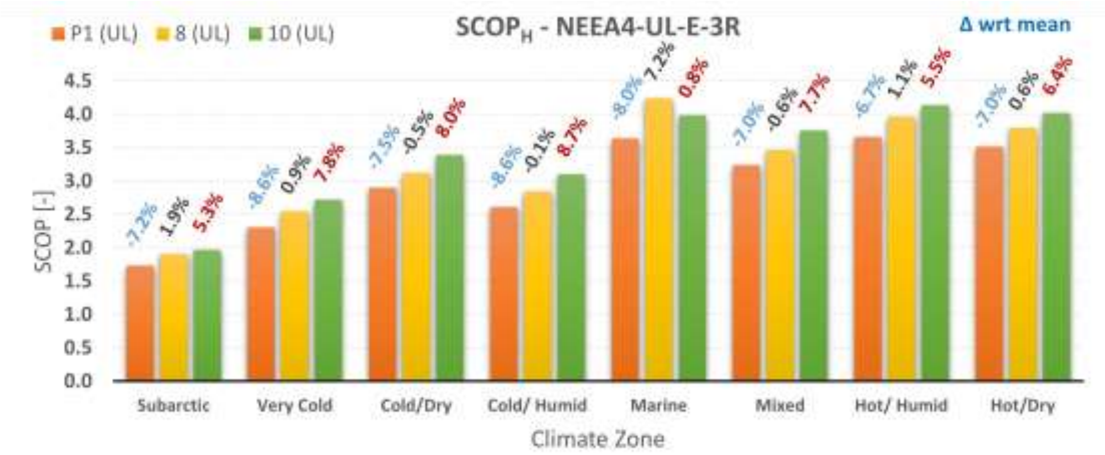


Figure 47. NEEA4 (UL) heating SCOP comparison in different climate zones for different tests

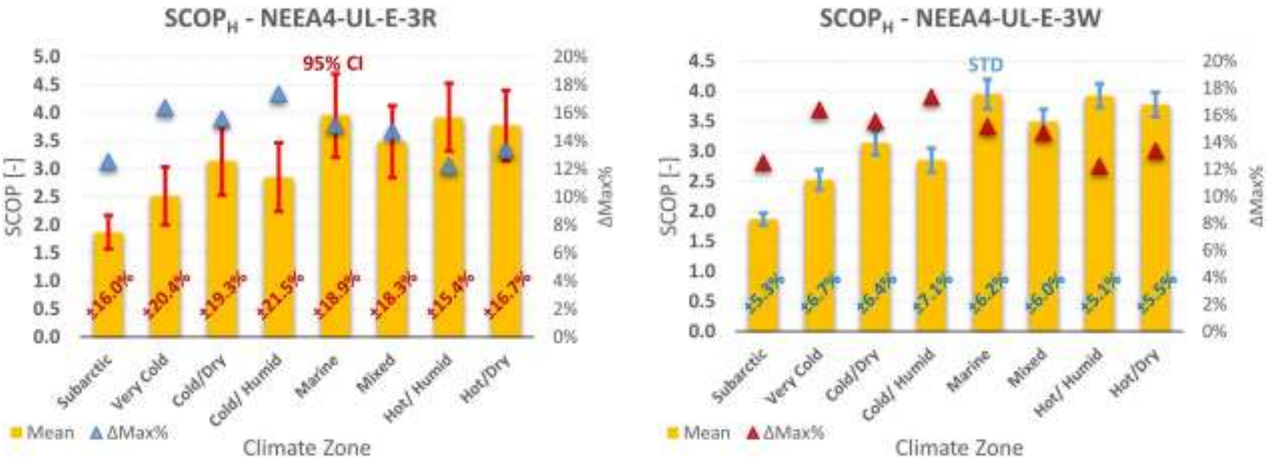


Figure 48. NEEA4 (UL) heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

Figure 49 shows the heating SCOP comparison for dataset NEEA4-UL-E-2W with test #8 and #10 only without considering test P1 which was performed around 16 months earlier than the other two. Without considering test P1, heating repeatability improves with the difference in SCOP varying from 3.3% to 8.5% and STD ±1.7% to ±4.3% across different climate zones, but a relatively larger difference compared to cooling for the same case.

Table 13 shows the overall SCOP repeatability results summary with the range and mean of three statistical parameters based on three repeated tests for NEEA4-UL-E-3R dataset. Overall, the repeatability was not good and comparatively worse in heating compared to cooling mode test results. This is due to differences in the test unit control behavior at the same test conditions for some conditions and also for some cases due to the need for improvements in the convergence criteria. On the other hand, the load-based testing approach identified situations where the equipment responded and performed differently to nearly the same load and ambient conditions. This indicates that the equipment behavior may depend on its history because of its embedded controls and highlights a strength of the load-best testing approach in identifying a control behavior that can occur in the field. Also, some possible areas of improvement were identified with convergence criteria and the testing approach to capture the test unit representative performance in a test interval concerning short on/off cycling and defrost behavior.

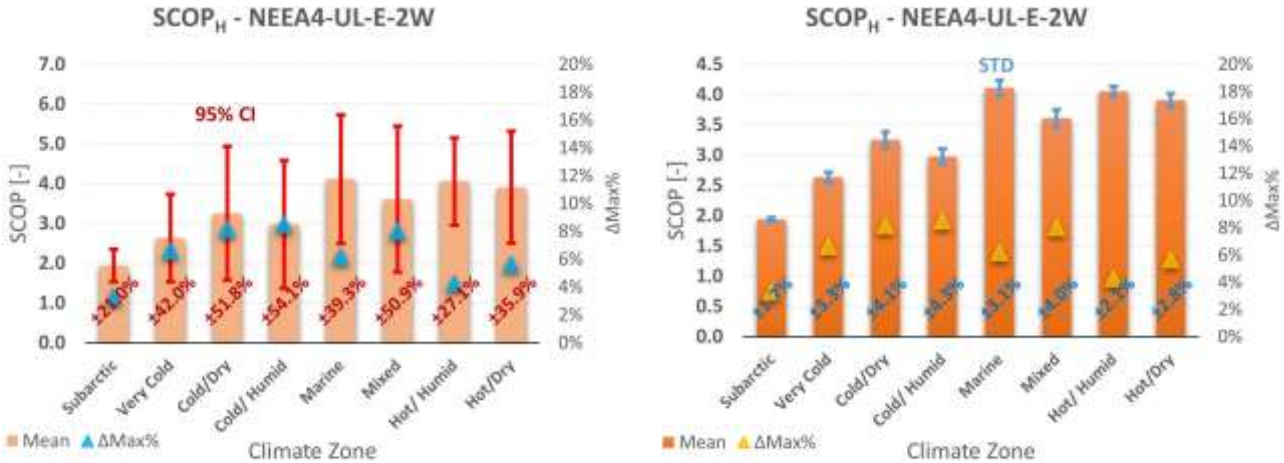


Figure 49. NEEA4 (UL) heating SCOP average and maximum difference for repeated tests without P1 test with a) 95% confidence interval, and b) standard deviation

Table 13. NEEA4-UL-E-3R dataset SCOP repeatability summary results

Metric	No. of Tests	∆SCOP (Max-Min) [%]			SCOP 95% CI [%]			SCOP STD [%]		
		Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
SCOP$_C$	3	1.5%	14.3%	9.3%	2.1%	20.4%	13.0%	0.7%	6.7%	4.3%
SCOP$_H$	3	12.2%	17.3%	14.6%	15.4%	21.5%	18.3%	5.1%	7.1%	6.0%

Test unit performance measured in test P1 was relatively different from the other two tests which could be due to various reasons discussed previously. Without considering test P1 in assessment with dataset NEEA4-UL-E-2W improves repeatability significantly for both cooling and heating as shown in Table 14. It should be noted that for heating SCOP, 95% confidence interval increased while differences decreased compared to Table 13 because it is also a strong function of the number of samples for student's t-distribution considered here, thus it is not a good statistical measure to compare variations with a different number of samples used in the analysis.

Table 14. NEEA4-UL-E-2W dataset SCOP repeatability summary results without test P1

Metric	No. of Tests	∆SCOP (Max-Min) [%]			SCOP 95% CI [%]			SCOP STD [%]		
		Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
SCOP$_C$	2	0.2%	1.6%	0.8%	1.1%	9.9%	5.1%	0.1%	0.8%	0.4%
SCOP$_H$	2	3.3%	8.5%	6.3%	21.0%	54.1%	40.3%	1.7%	4.3%	3.2%

4.3 NEEA4 (PG&E)

In this section, EXP07 repeatability assessment analysis results are presented for the NEEA4 unit, a 1-ton mini-split ductless variable-speed heat pump system, based on two sets of test results from the PG&E lab with dataset NEEA4-PGE-E-2W.

4.3.1 Cooling Mode

Figure 50 and Figure 51 show the COP comparison of cooling dry coil and humid coil test results, respectively, with unit operation mode during the convergence period and possible issues with test interval for two repeated tests in dataset NEEA4-PGE-E-2W. Similar unit operation mode was captured during convergence period at same outdoor temperature test conditions except dry coil test interval at 86°F ambient temperature. Figure 52 shows the mean COP of repeated tests for dry and humid coil test intervals with STD and the absolute COP percentage difference as a statistical measure of repeatability. The difference in COP between two tests at the same test conditions varied from 0.5% to 2.4% and STD from ±0.3% to ±1.2% for dry coil test intervals. Similar results were observed for humid coil test intervals, with COP difference varying from 0.8% to 2.1% and STD from ±0.4% to ±1%. Overall test unit showed good repeatability in cooling mode performance at PG&E.

However, some possible issues related to convergence criteria and testing approach were observed. For example, in test #8 dry coil test interval at 86°F outdoor temperature, even though the unit was cycling on/off, the convergence was found in steady operation in the beginning as shown in Figure 53, thus not capturing the representative performance including on/off cycles.

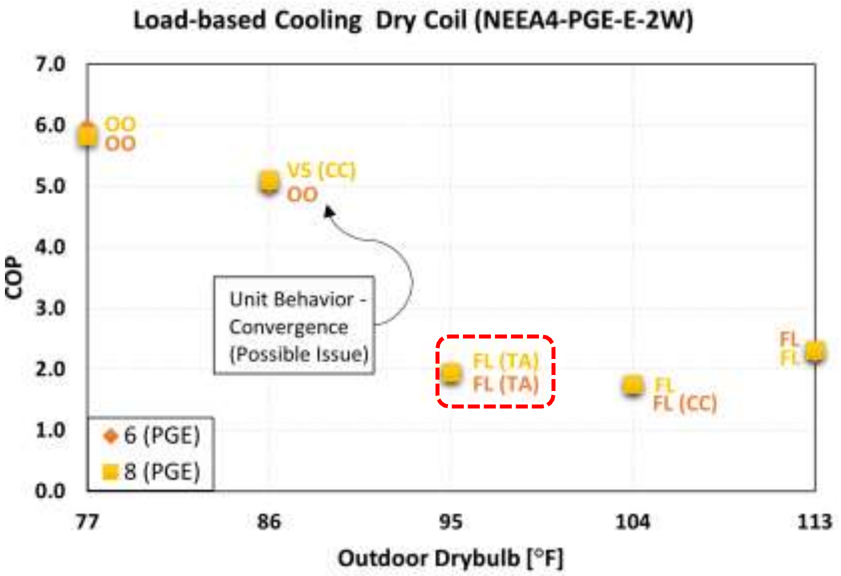


Figure 50. NEEA4 (PG&E) cooling dry coil test intervals COP for different tests with unit behavior during convergence

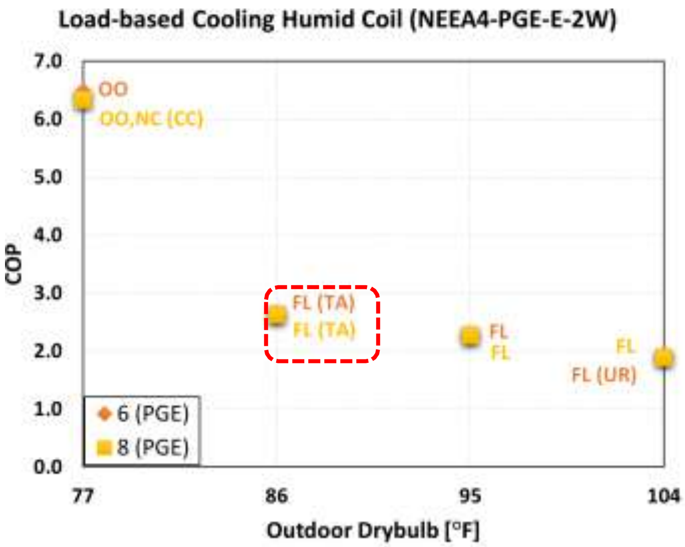


Figure 51. NEEA4 (PG&E) cooling humid coil test intervals COP for different tests with unit behavior during convergence

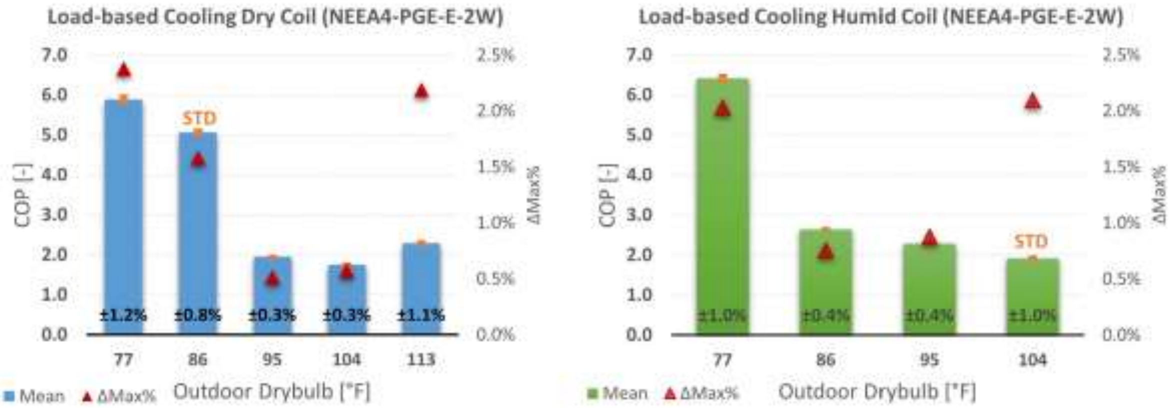


Figure 52. NEEA4 (PG&E) cooling dry coil and humid coil test intervals mean COP with STD and maximum difference among repeated tests

Further, in dry coil test interval CC at 95°F outdoor temperature, a full-load test was performed, but looking at the cooling rate and building load during convergence in Figure 54, it can be seen that the unit did not run out of capacity at this condition and a load-based test should have been performed to measure more representative performance for this test condition. However, in the beginning, it seemed like the unit was not able to match the building load due to its slow ramp-up and the indoor temperature increased above the limit (thermostat setpoint + 2°F) to perform the full-load test, which led to the decision to change to full-load test. This issue could be resolved by updating the testing approach for cases like this and allowing the unit a longer duration to match the building load in the beginning. Also, it would be good to have a provision in the testing approach to double-check whether the decision to change to a load-based test makes sense. A similar issue was observed in the humid coil test interval at 86°F outdoor temperature (CD) where a full-load test was also performed. In test #8 humid coil test interval at 77°F ambient temperature (CE), the test unit cycled on/off, and convergence was not found. However, the convergence period did not include complete on/off cycles showing a possible issue with convergence criteria implementation.

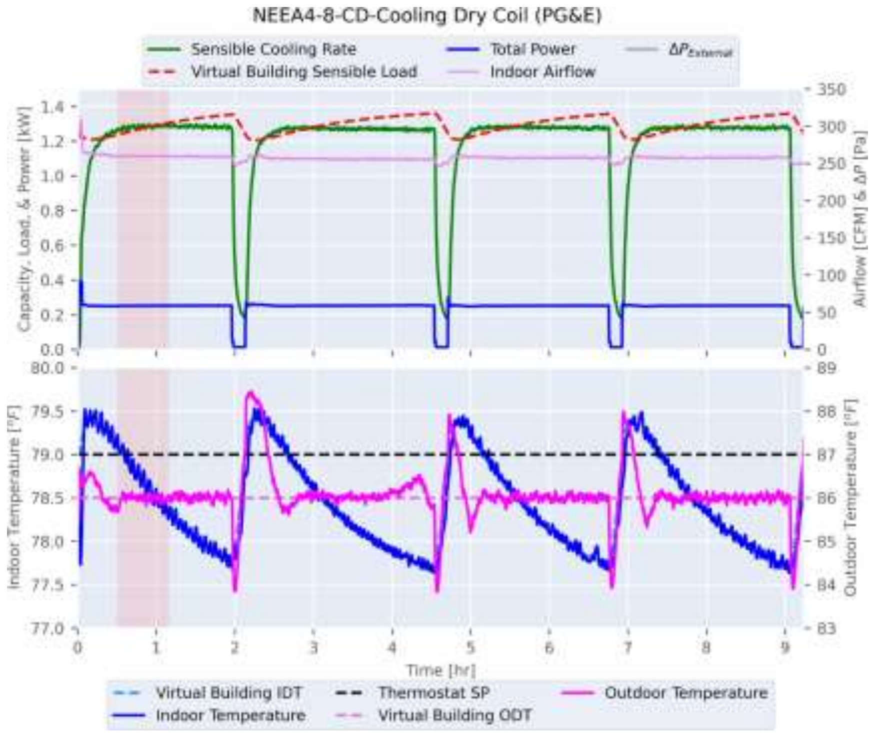


Figure 53. NEEA4-8 (PG&E) performance and temperature variations for cooling dry coil test interval CD

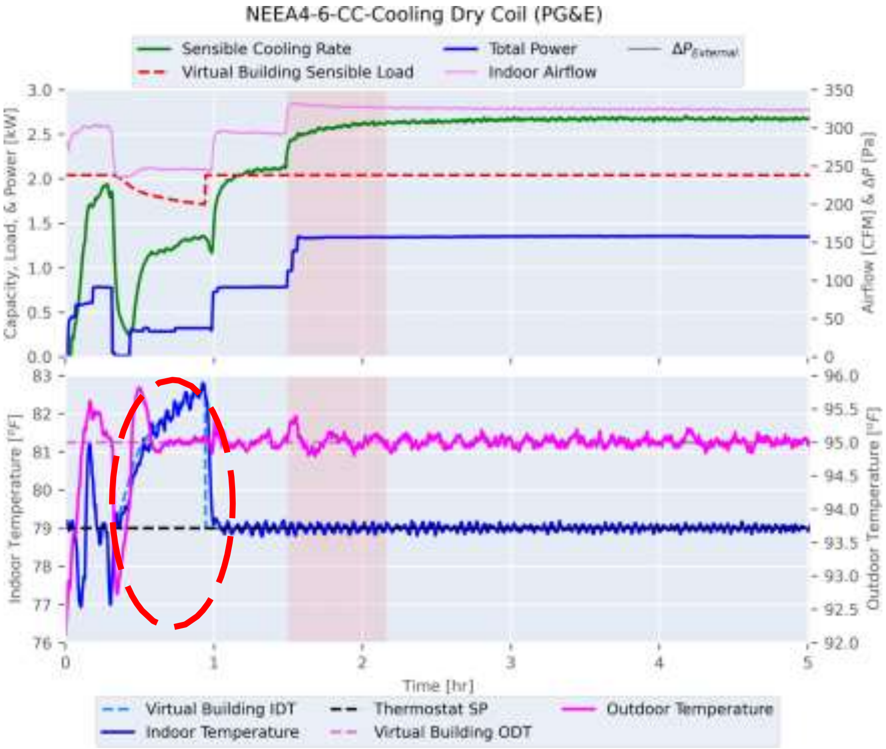


Figure 54. NEEA4-6 (PG&E) performance and temperature variations for cooling dry coil test interval CC

Figure 55 and Figure 56 show the comparison of NEEA4 estimated cooling SCOP for different climate zones based on measured performance in two sets of different tests at PG&E in dataset NEEA4-PGE-E-2W. Overall good repeatability is observed in cooling SCOP. For climate zones that utilize cooling dry coil test results in SCOP estimation (3 out of 7), the maximum difference varies from 0% to 0.2% and STD from ±0% to ±0.1%. For climate zones utilizing humid coil test results, a relatively larger variation in cooling SCOP was observed with a maximum difference of around 1.2% and STD around ±0.6%.

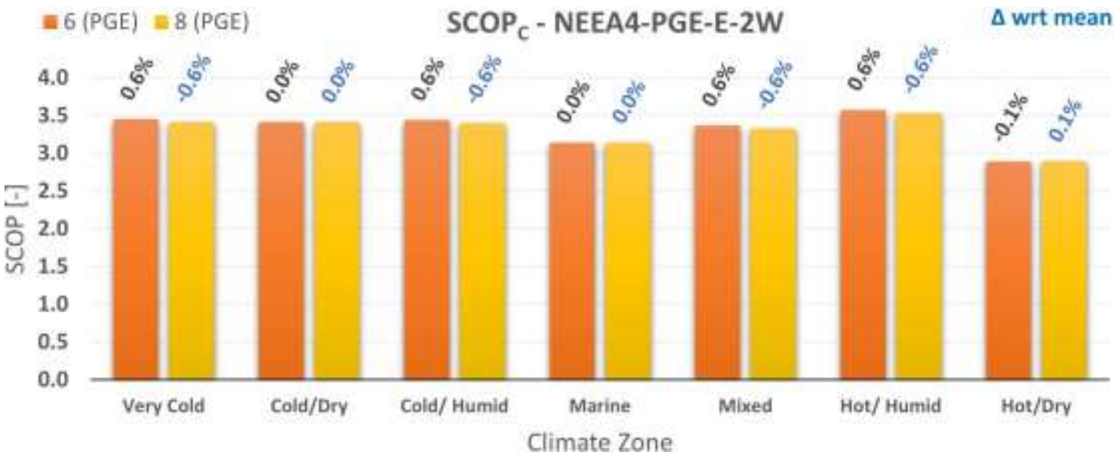


Figure 55. NEEA4 (PG&E) cooling SCOP comparison in different climate zones for different tests

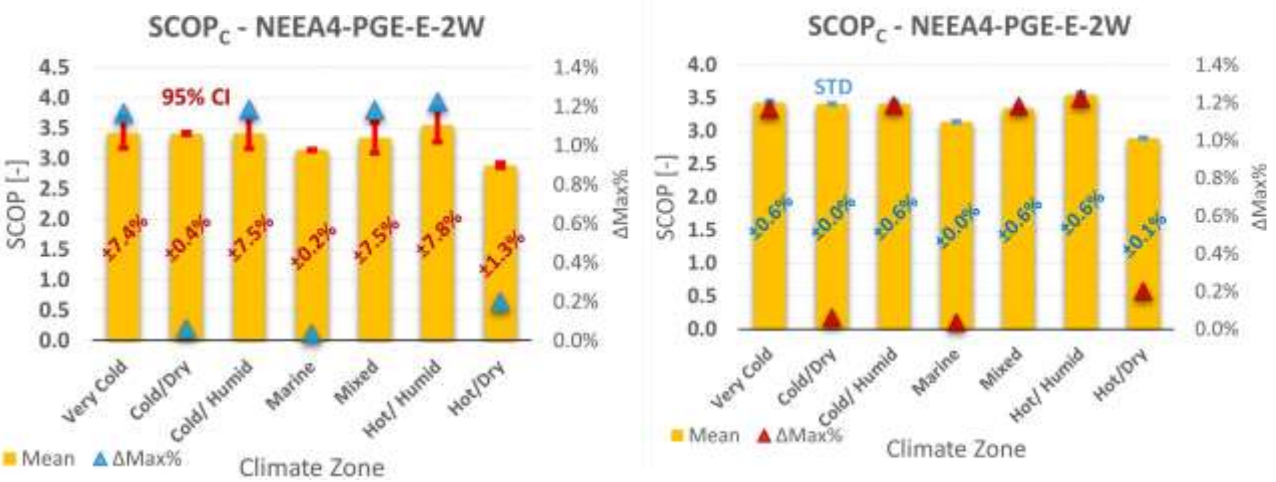


Figure 56. NEEA4 (PG&E) cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

4.3.2 Heating Mode

Figure 57 and Figure 58 show the comparison of NEEA4 performance in heating continental test intervals for 2 repeated sets of tests in dataset NEEA4-PGE-E-2W. At PG&E, unlike UL, test unit performance was not measured at the lowest outdoor temperature (-10°F) test interval due to test facility limitations. At the same test conditions, similar unit operation modes were observed during convergence in two sets of tests, with an absolute difference in COP varying from 0.2% to 12.2% and STD from ±0.1% to ±6.1%, with larger variations compared to cooling test results. Relatively large differences in performance were observed for test intervals at 47°F, 17°F, and 5°F ambient temperatures, two of which are full-load tests. Similar to heating test results at UL, in almost all the test intervals some possible issues were noted related to either convergence criteria, testing approach, or inconsistent unit dynamic response. The convergence criteria issue is mainly related to its implementation to not capture the complete cycles during unit defrost cycling behavior with non-convergence in test interval. The testing approach's possible issues are related to improper thermostat offset settings and the identification approach of the full-load test changeover.

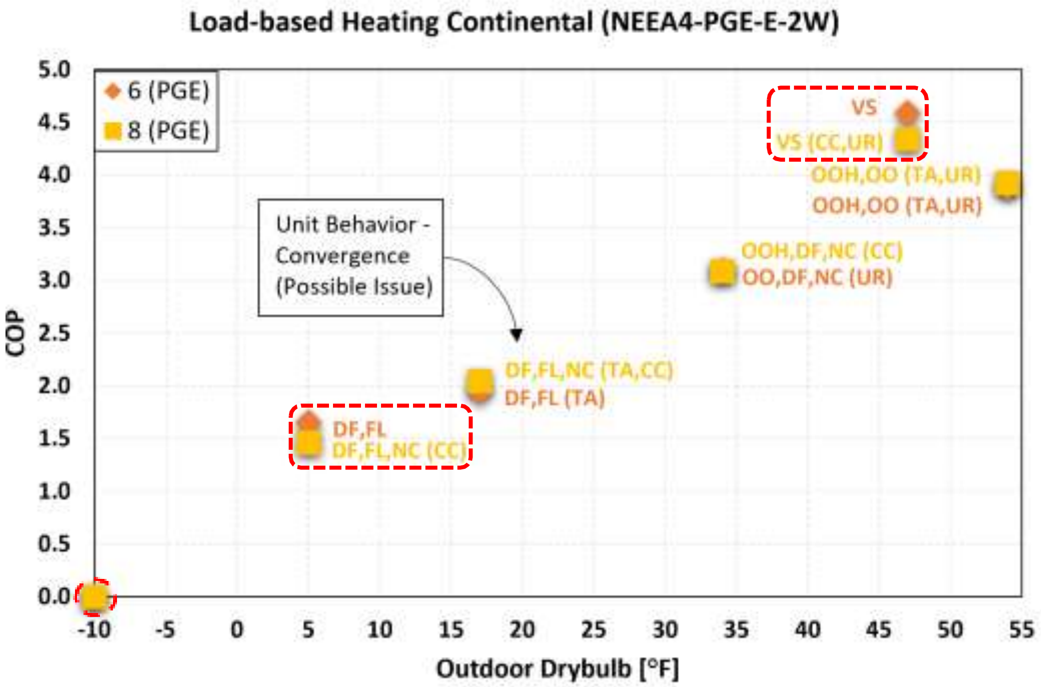


Figure 57. NEEA4 (PG&E) heating continental test intervals COP for different tests with unit behavior during convergence

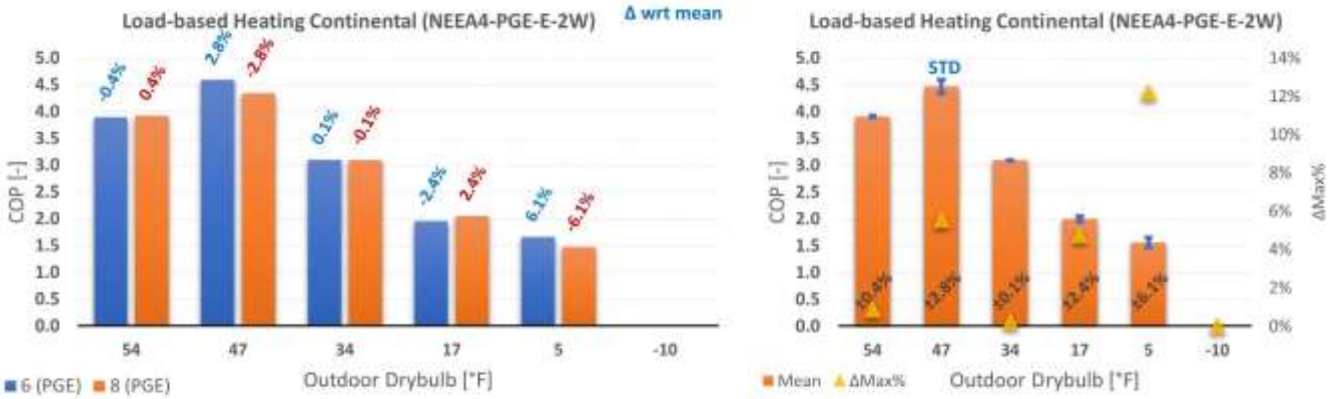


Figure 58. NEEA4 (PG&E) heating continental test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

For the 54°F test interval (HF) as shown in Figure 59 for test 6, there is a possible issue with the testing approach as the indoor temperature was maintained at around 77.7°F, quite high compared to the thermostat setpoint of 70°F, maybe a result of an incorrect thermostat offset setting. Also, a mix of short and long on/off cycling behavior was observed similar to UL results, which could be attributed to inconsistent unit dynamic response.

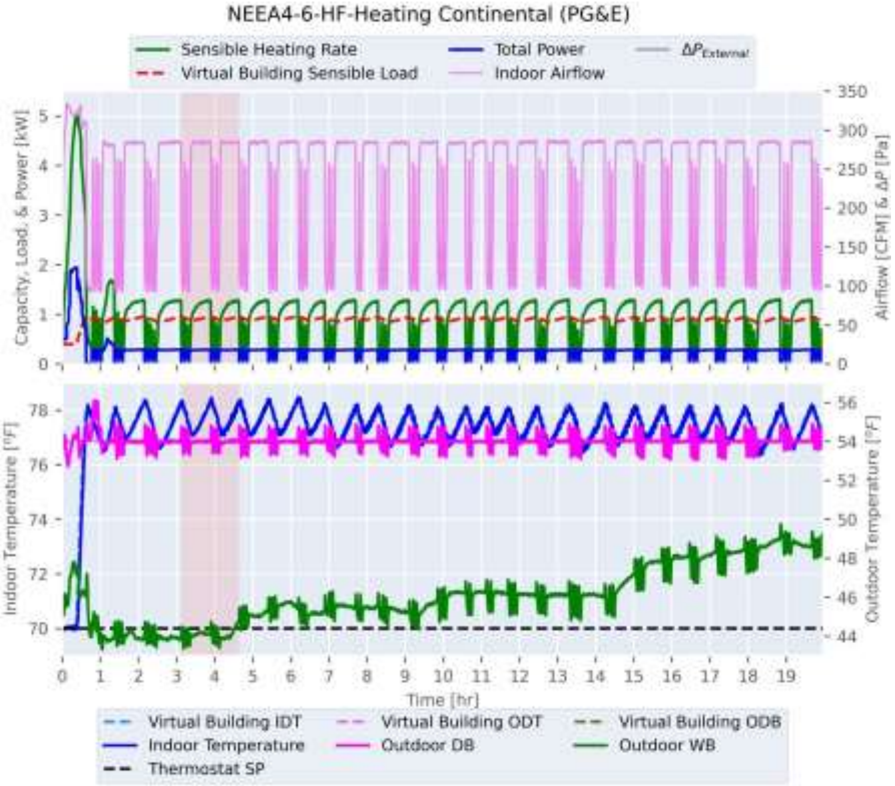


Figure 59. NEEA4-6 (PG&E) performance and temperature variations for heating continental test interval - HF (54°F)

In 47°F outdoor temperature test interval, around 5.6% lower COP was measured in test #8 compared to test #6. In both tests, the unit showed similar behavior with cycling on/off in the beginning and then operating in steady variable speed mode where convergence was found as shown in Figure 60 for test #8. The difference in performance could be due to variation in test unit dynamic response at the same test condition. In this test interval, indoor temperature was also maintained around 5°F higher than the thermostat setpoint, similar to the 54°F test interval, showcasing a possible issue with the testing approach related to the thermostat offset setting. At 34°F ambient temperature, the unit showed a mix of on/off and defrost cycles together, but the unit did not reach a steady periodic response, and convergence was not achieved in both tests as shown for test #6 in Figure 61. This could be due to the test unit's inconsistent dynamic response. Also, in the final converged window, complete on/off cycles were not taken to calculate average performance for the test interval as per convergence criteria, showing a possible issue with convergence criteria interpretation or implementation.

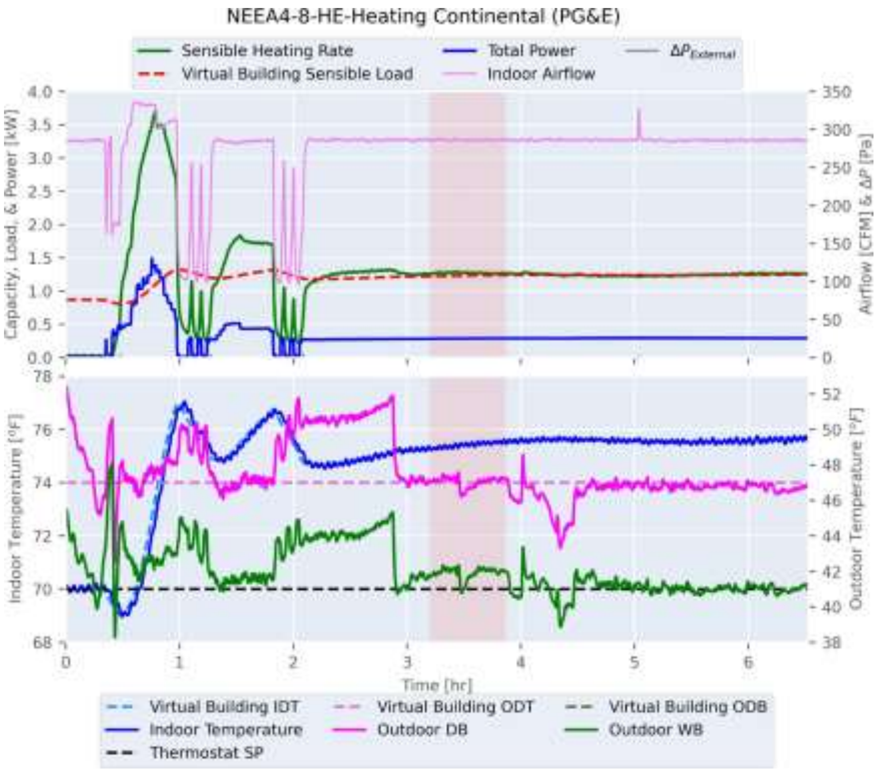


Figure 60. NEEA4-8 (PG&E) performance and temperature variations for heating continental test interval HE (47℉)

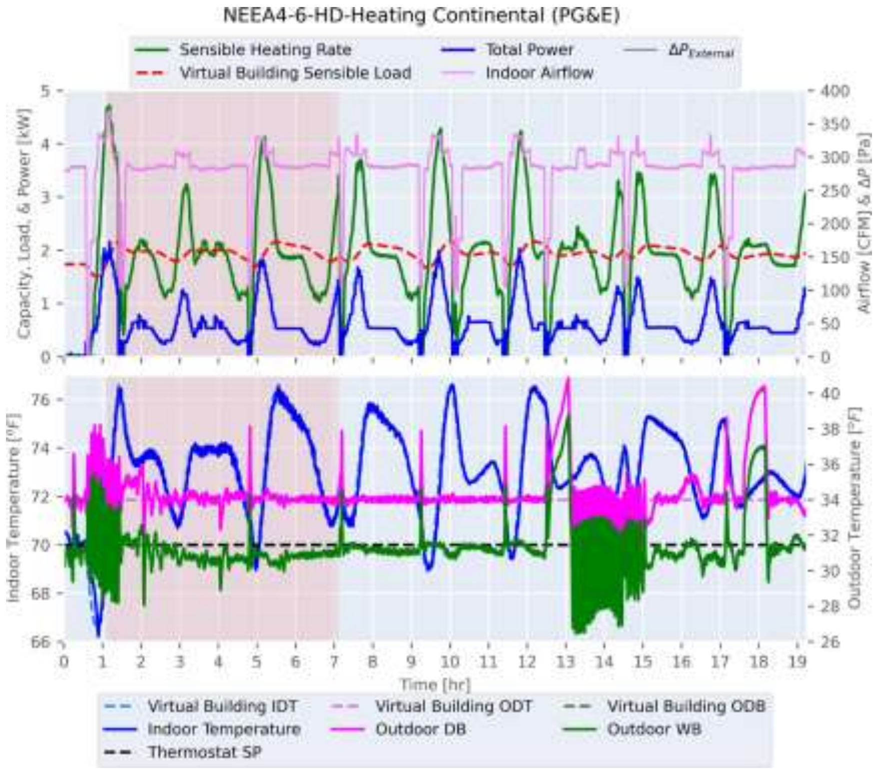


Figure 61. NEEA4-6 (PG&E) performance and temperature variations for heating continental test interval HD (34℉)

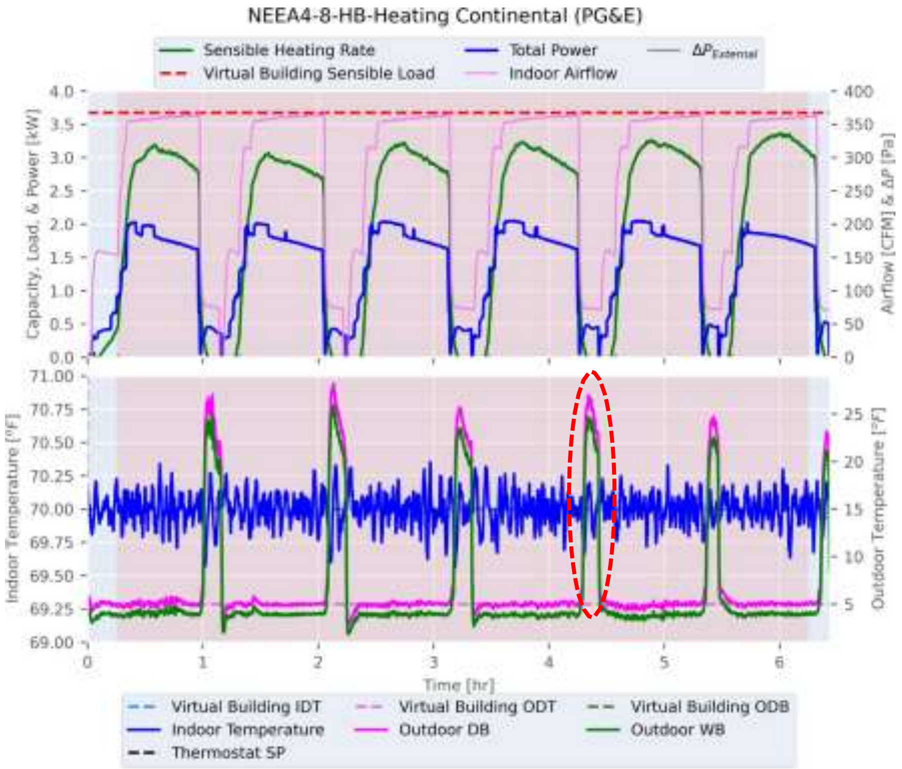


Figure 62. NEEA4-8 (PG&E) performance and temperature variations for heating continental test interval HB (5℉)

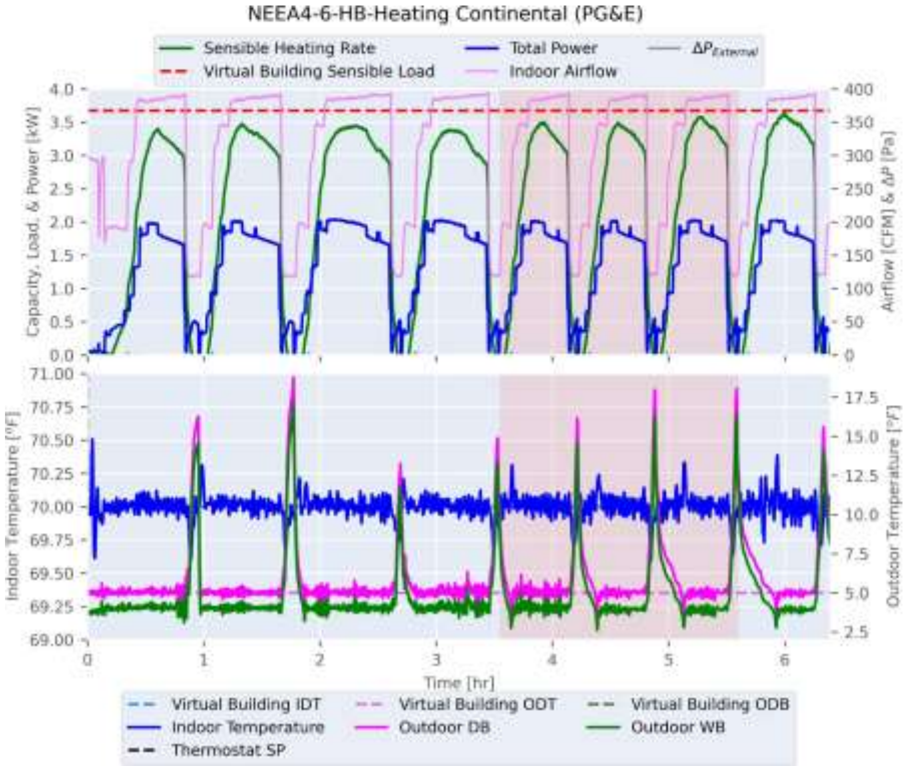


Figure 63. NEEA4-6 (PG&E) performance and temperature variations for heating continental test interval HB (5℉)

At 17°F and 5°F outdoor temperature, full-load tests were performed where the unit went into defrost cycles. In these test intervals, possible convergence criteria issues related to not taking complete on/off cycles in average performance and also testing approach to decide full-load test similar to cooling results as discussed above were observed. Figure 62 shows the test unit performance with time at 5°F ambient temperature in test #8, where around 12.2% lower COP was measured compared to test #6 at the same conditions as shown in Figure 63. Compared to test #6, longer defrosts were observed in test #8, and convergence was not achieved in successive defrost cycles, overall resulting in lower COP. It is also interesting to note that during defrosts, large fluctuations in outdoor temperature occur which would also affect the test results' repeatability and reproducibility, therefore test facility control is also critical here. These large fluctuations in outdoor temperature do not satisfy the test operating tolerance as per EXP07, which might make this test void. A further investigation is required to define the appropriate test condition and operating tolerances in EXP07 which are large enough for the current test facilities' capabilities and small enough to ensure repeatability and reproducibility within certain bounds.

Figure 64 and Figure 65 show the repeatability comparison of test unit performance in heating marine test intervals. Overall reasonable repeatability was noted with an absolute difference in COP varying from 1.1% to 4.4% and STD ±0.5% to ±2.2%, similar to continental test intervals at the same ambient temperatures. In heating marine test intervals, similar possible issues related to the convergence criteria, testing approach, and inconsistent unit dynamic repose were observed as described above for continental conditions.

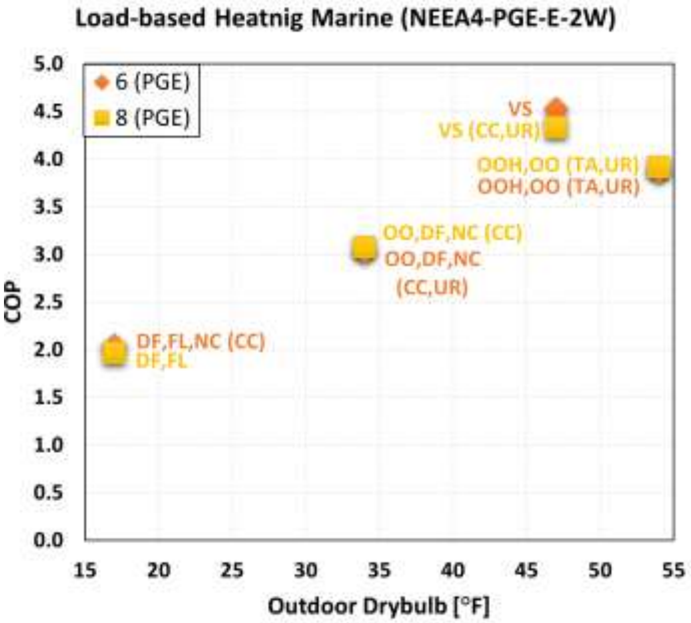


Figure 64. NEEA4 (PG&E) heating marine test intervals COP for different tests with unit behavior during convergence

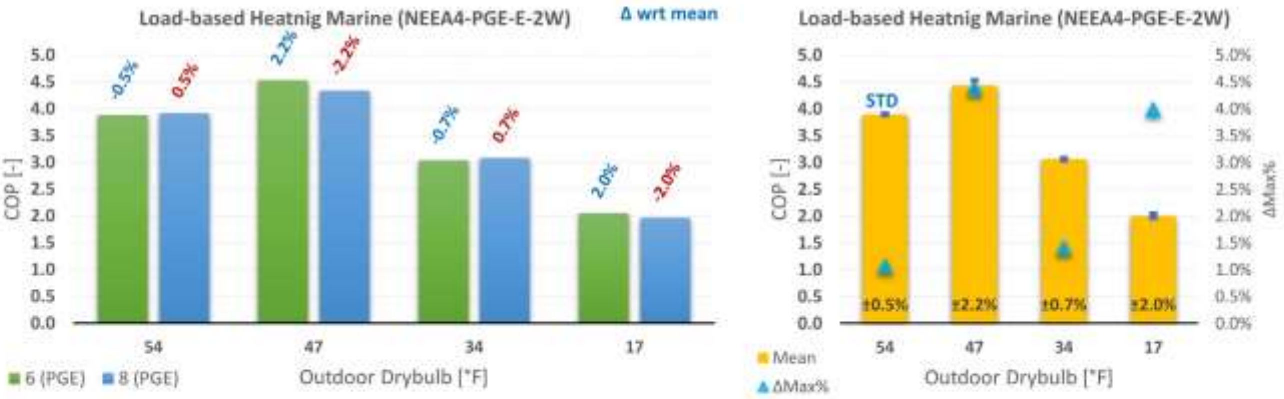


Figure 65. NEEA4 (PG&E) heating marine test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Figure 66 and Figure 67 show comparisons of estimated heating SCOP for different climate zones along with the maximum differences and average performances with a 95% CI and STD based on two sets of tests at PG&E. Overall good repeatability was noted with differences in heating SCOP varying from 0.4% to 3.2% and STD from ±0.2% to ±1.6%. Similar to cooling, unit performance in test #8 was relatively low compared to test #6 for heating mode.

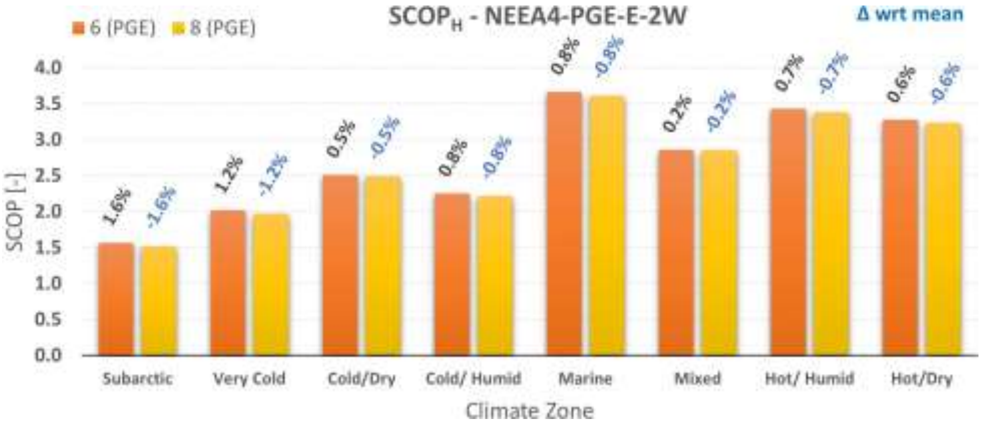


Figure 66. NEEA4 (PG&E) heating SCOP comparison in different climate zones for different tests

Table 15 shows the overall SCOP repeatability results summary with the range and mean of three statistical parameters for NEEA4-PGE-E-2W dataset with two repeated tests. Overall, the repeatability was good and comparatively better in cooling compared to heating mode test results. Also, some possible areas of improvement were identified with convergence criteria and a testing approach to capture the test unit representative performance in a test interval as discussed above.

Table 15. NEEA4-PGE-E-2W dataset SCOP repeatability summary results

Metric	No. of Tests	ΔSCOP (Max-Min) [%]			SCOP 95% CI [%]			SCOP STD [%]		
		Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
SCOP$_C$	2	0.0%	1.2%	0.7%	0.2%	7.8%	4.6%	0.0%	0.6%	0.4%
SCOP$_H$	2	0.4%	3.2%	1.6%	2.6%	20.3%	10.1%	0.2%	1.6%	0.8%

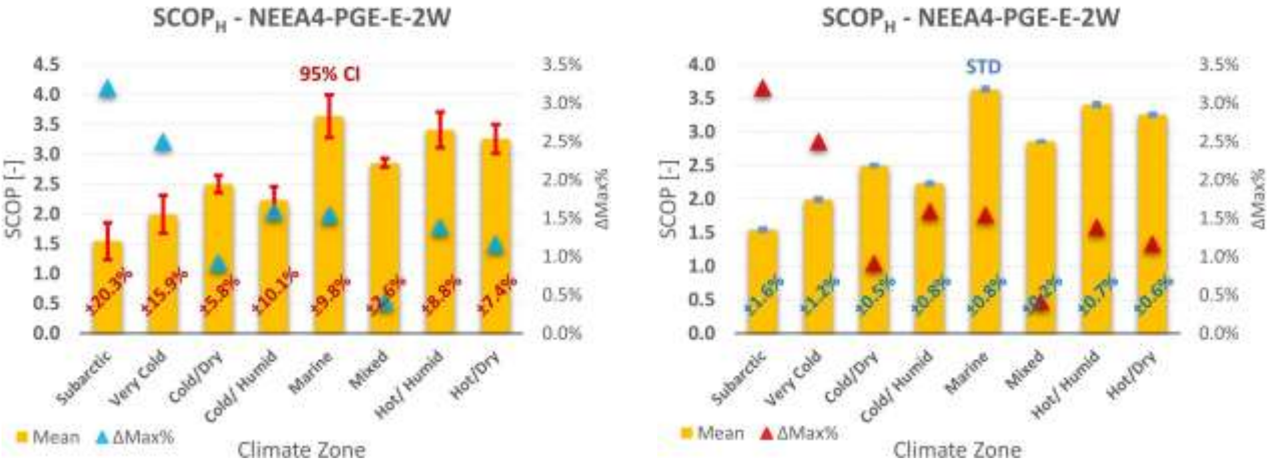


Figure 67. NEEA4 (PG&E) heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

4.4 NEEA7(B) (UL)

In this section, EXP07 repeatability assessment analysis results are presented for the NEEA7(B) (NEEA7 & NEEA7B), a 3-ton split-type variable-speed heat pump system, based on three sets of test results from UL for two datasets, NEEA7(B)-UL-E-3R and NEEA7-UL-E-2W.

4.4.1 Cooling Mode

Figure 68 and Figure 69 show the comparison of UUT performance and unit operation mode during convergence for cooling dry coil test intervals for the three sets of tests in dataset NEEA7(B)-UL-E-3R. It should be noted that tests P2 and P3 were performed back-to-back with NEEA7, whereas test #11 was performed around 22 months later after reinstalling the same model of the test unit (NEEA7B) at UL after testing it first at PG&E. Nonetheless, at same test conditions similar operation mode behavior was observed during convergence. Generally, test unit performance was measured highest in test P2 and lowest in test #11. The maximum difference in COP among three tests across different test intervals varied from 3.4% to 12.9% and STD ±1.4% to ±5.4%, showing poor repeatability in some test intervals.

The largest differences were observed in test intervals at 77°F and 86°F ambient temperature where the unit was cycling on/off and also in the test interval at 104°F outdoor temperature with variable speed steady and hybrid operation mode. The difference in COP during on/off cycling test intervals was possibly due to differences in test unit dynamic response i.e., compressor speed ramp up during cycle on period and indoor temperature maintained. This is evident in Figure 70 and Figure 71, which show the test unit performance in 86°F test interval (CD) for tests P2 and #11, respectively. In test #11, the average cooling rate during convergence was around 2.1% lower than test P2, but power consumption was around 7.2% higher, resulting in around 10.5% lower COP. The effect of test unit dynamic response could be seen in the sensible cooling rate as well as indoor temperature variation between the two tests. In 104°F outdoor temperature test interval, even though similar test unit dynamic behavior was observed in different tests, the differences in COP resulted from the period during which convergence was found as shown in Figure 72 and Figure 73 for tests P2 and #11, respectively. The other possible causes for the difference observed could be due to the test step after re-installation, charge, measurement, and test uncertainty.

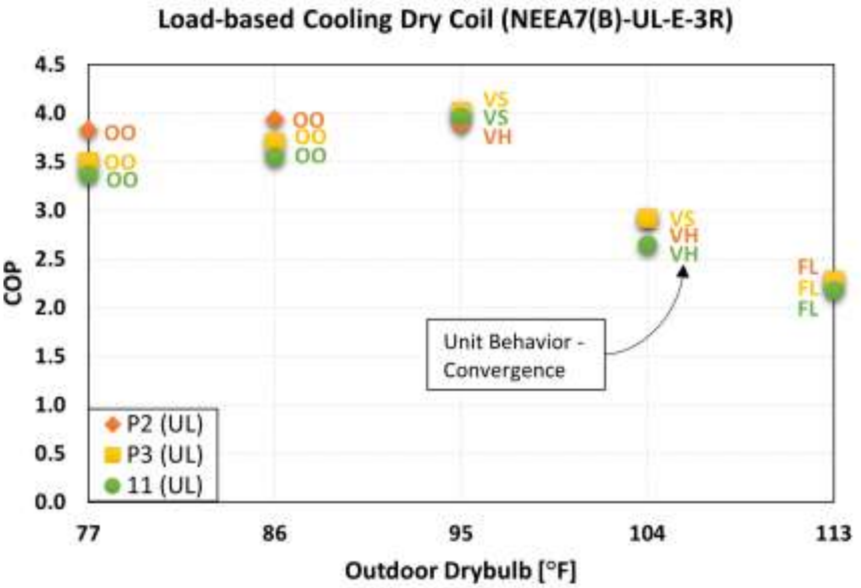


Figure 68. NEEA7(B) (UL) cooling dry coil test intervals COP for different tests with unit behavior during convergence

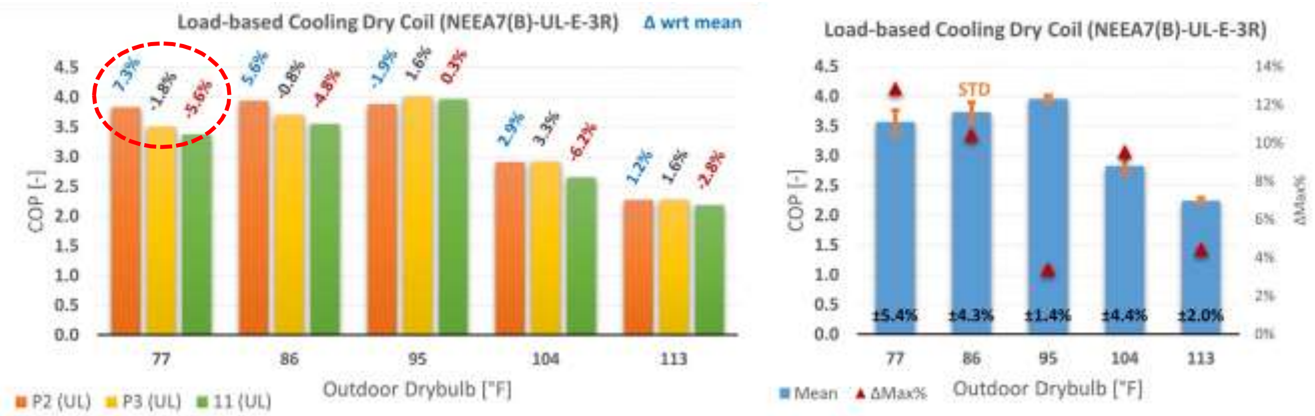


Figure 69. NEEA7(B) (UL) cooling dry coil test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

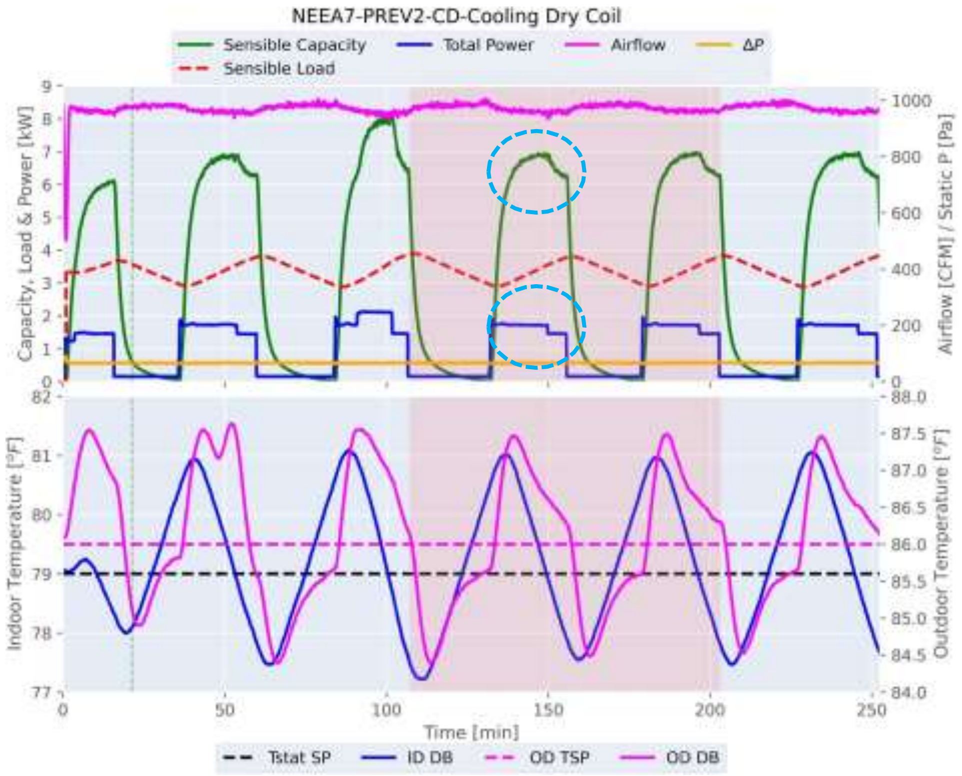


Figure 70. NEEA7-P2 (UL) performance and temperature variations for cooling dry coil test interval CD (86°F)

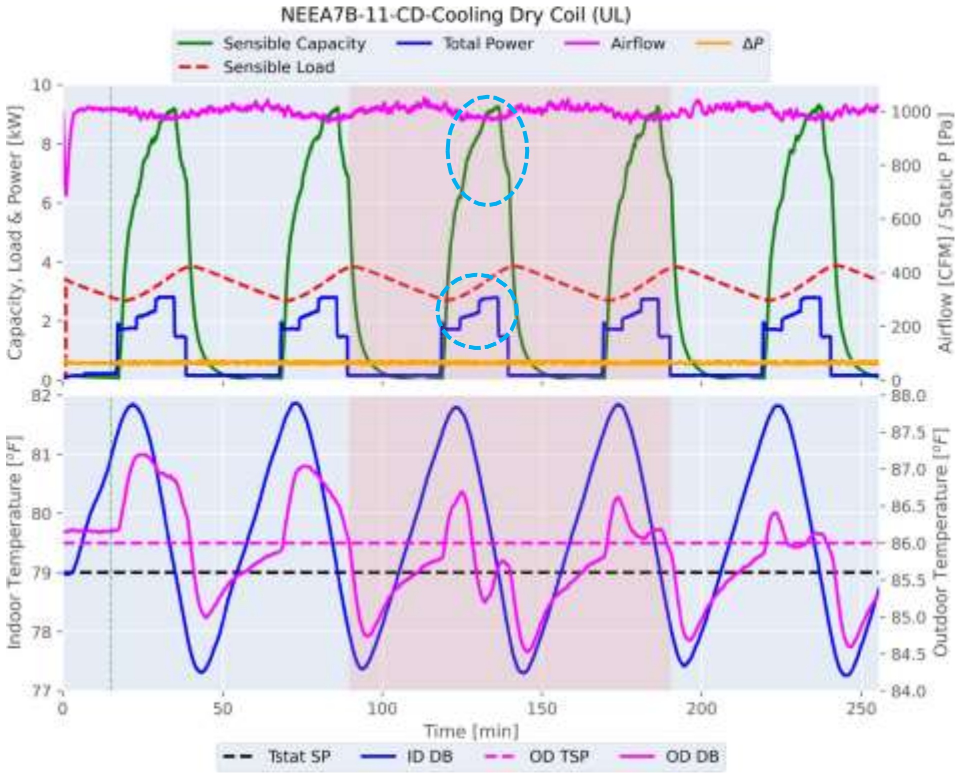


Figure 71. NEEA7B-11 (UL) performance and temperature variations for cooling dry coil test interval CD (86°F)

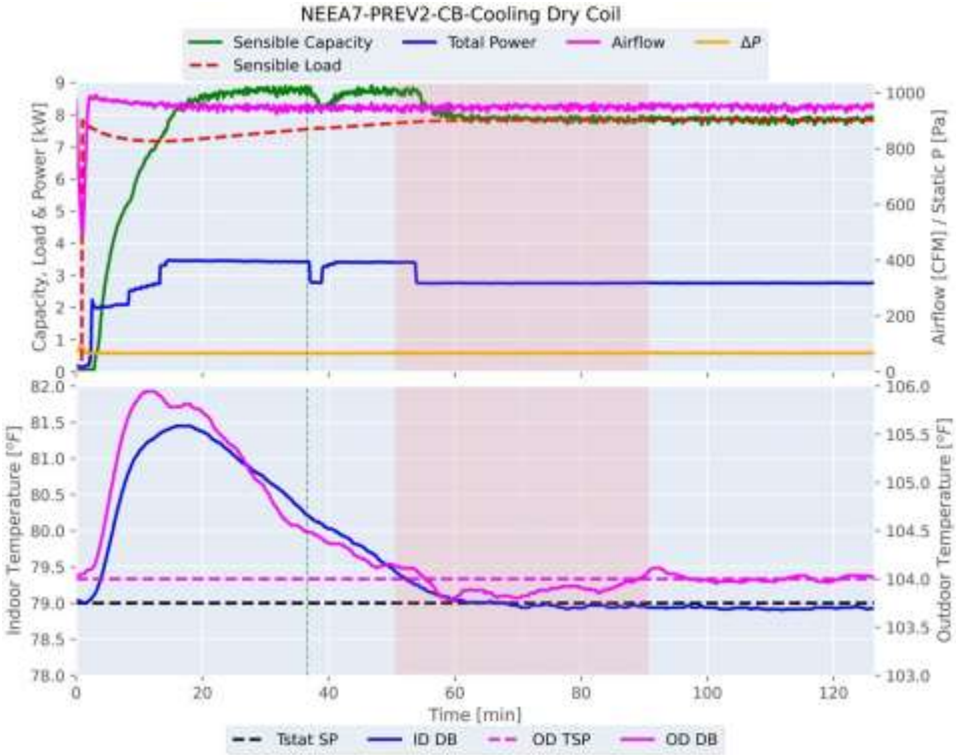


Figure 72. NEEA7-P2 (UL) performance and temperature variations for cooling dry coil test interval CB (104℉)

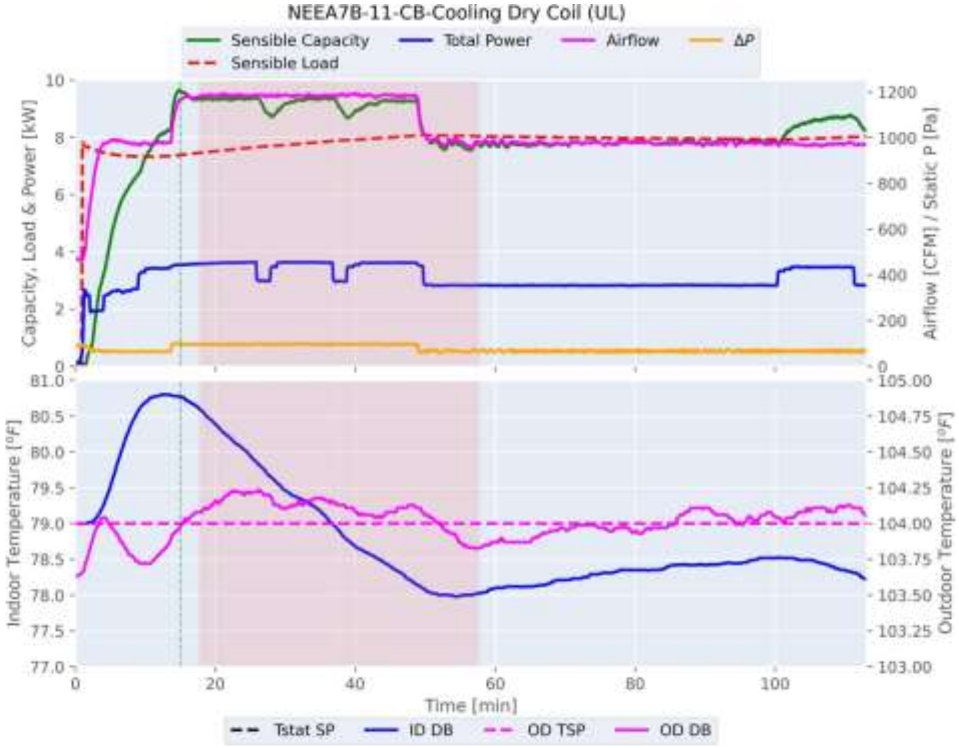


Figure 73. NEEA7B-11 (UL) performance and temperature variations for cooling dry coil test interval CB (104℉)

Figure 74 and Figure 75 show the performance comparison of cooling humid coil test intervals. Overall good repeatability was observed except for 95°F outdoor condition test interval. The maximum difference in COP between three sets of tests varied from 2% to 24.1% and STD from ±0.9% to ±10.4% across different test intervals. In the 95°F ambient temperature test interval, the difference in COP is mainly driven by the test unit's higher COP in test P2 compared to tests P3 and #11. This is primarily due to the difference in test unit compressor loading captured during the convergence period as shown in Figure 76 and Figure 77. In both cases, the convergence period does not seem to capture test unit representative performance including low and high compressor speed modes. It seems that the unit was loading and unloading with a somewhat regular pattern. In this case, it might make sense from a repeatability and representativeness point of view to consider a complete cycle, and correspondingly the convergence criteria should be updated and tested.

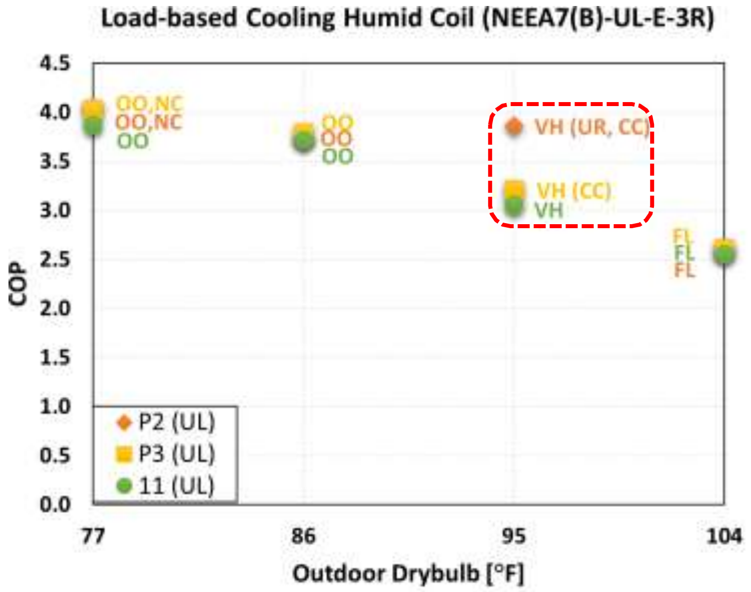


Figure 74. NEEA7(B) (UL) cooling humid coil test intervals COP for different tests with unit behavior during convergence

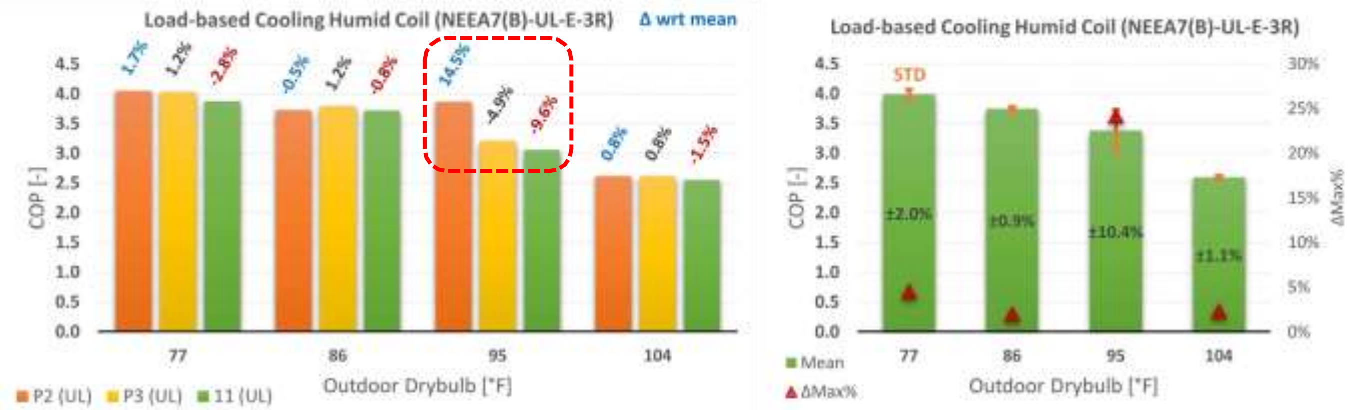


Figure 75. NEEA7(B) (UL) cooling humid coil test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

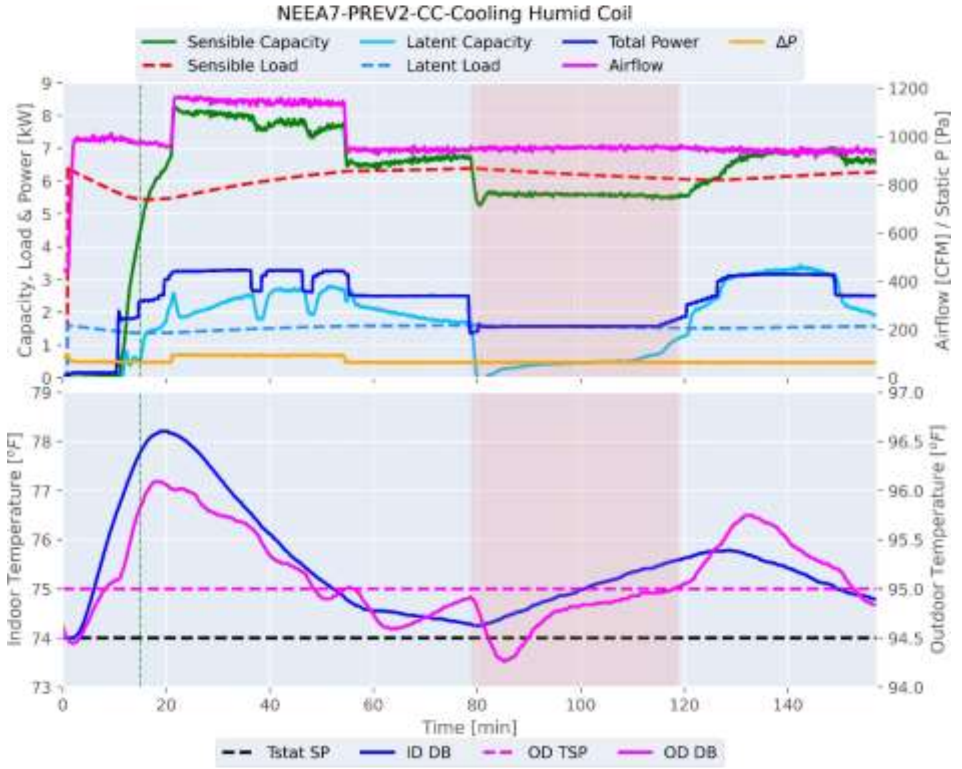


Figure 76. NEEA7-P2 (UL) performance and temperature variations for cooling humid coil test interval – CC (95℉)

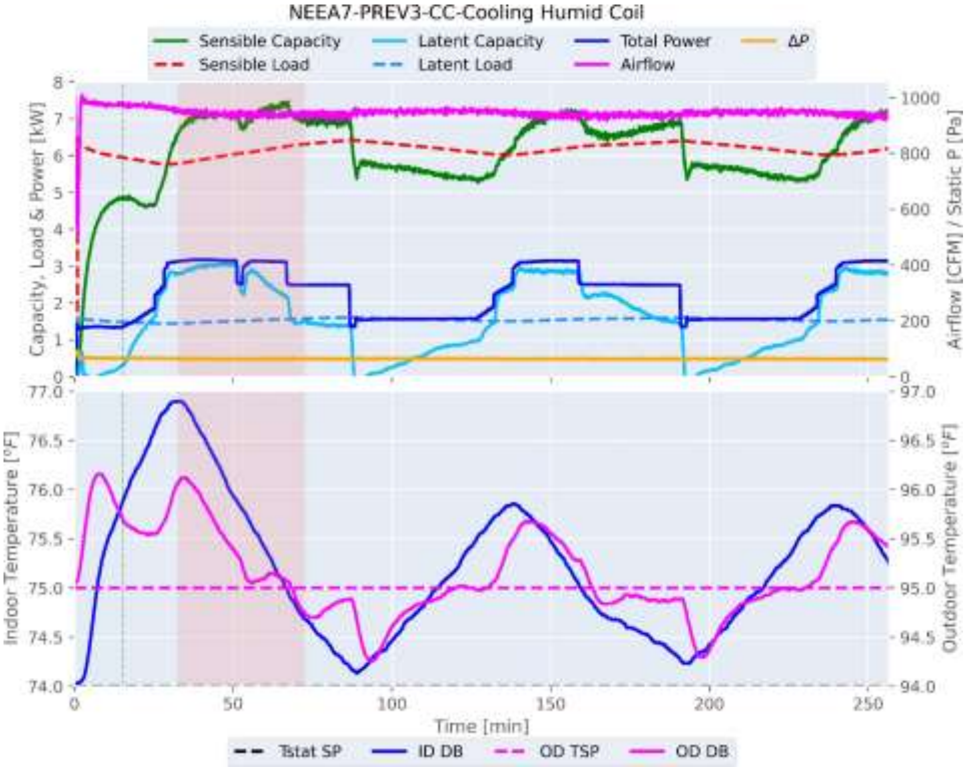


Figure 77. NEEA7-P3 (UL) performance and temperature variations for cooling humid coil test interval - CC (95℉)

Figure 78 and Figure 79 show a comparison of NEEA7(B) estimated cooling SCOP based on three sets of test results in dataset NEEA7(B)-UL-E-3R at UL for different climate zones. Overall reasonable repeatability was noted, and similar differences were observed across different climate zones with slightly better results in climate zones utilizing dry coil tests corresponding to the COP repeatability presented above. The maximum difference in SCOP across various climate zones between repeated tests varies from 5.1% to 6.8% and STD from ±2.1% to ±2.8%.

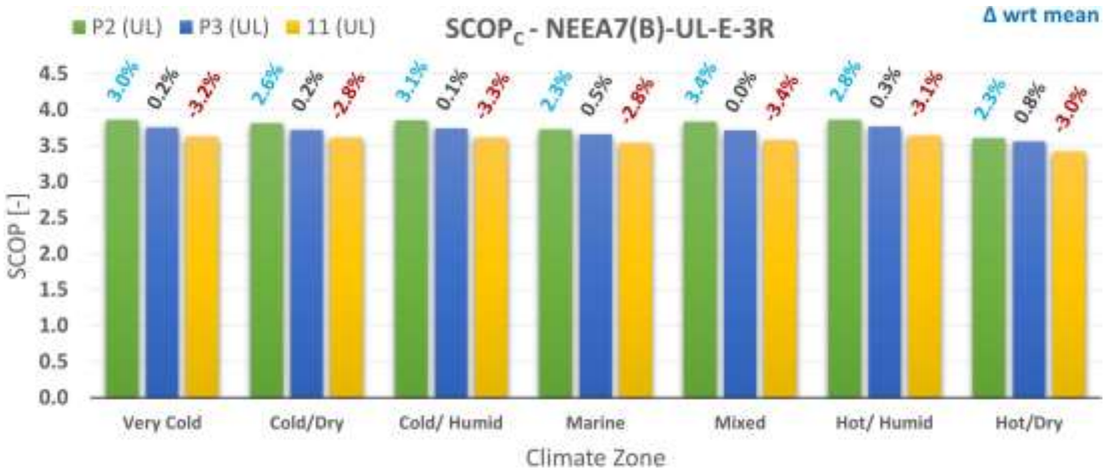


Figure 78. NEEA7(B) (UL) cooling SCOP comparison in different climate zones for different tests

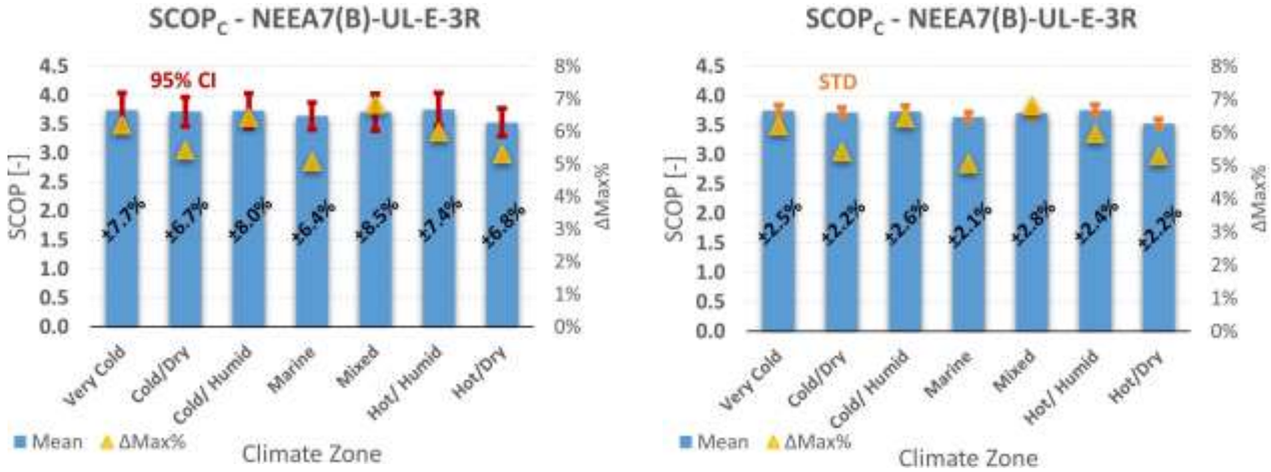


Figure 79. NEEA7(B) (UL) cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

4.4.2 Heating Mode

Figure 80 and Figure 81 show the comparison of NEEA7(B) heating continental test intervals performance for 3 sets of tests at UL in dataset NEEA7(B)-UL-E-3R. At the same test conditions, in general, the unit showed a similar dynamic response in different tests; however, in some test intervals, a similar unit operation mode was not captured during the converged period at the same test conditions. Also, some possible issues related to convergence criteria, testing approach, and inconsistent unit dynamic response were identified for some test intervals as shown in Figure 80. Here, the convergence criteria issues mainly related to its not capturing the overall representative performance during; i) unit on/off cycling with defrost and test interval non-convergence; ii) unit utilizing defrost or cycle on/off and convergence criteria didn’t include that in overall test interval performance. The testing approach issue is related to the case when UUT was going into defrost and the test was not run long enough to capture enough complete defrost cycles to capture overall representative performance. Across different test intervals, the maximum difference in COP among 3 tests varies from 2.5% to 14.3% and STD from ±1.1% to ±6.6%. Comparatively, a large variation was observed in measured performance for 17°F, 54°F, and 5°F outdoor temperature test intervals. Best repeatability was noted at -10°F ambient temperature full-load test interval with the maximum difference in COP of around 2.5%.

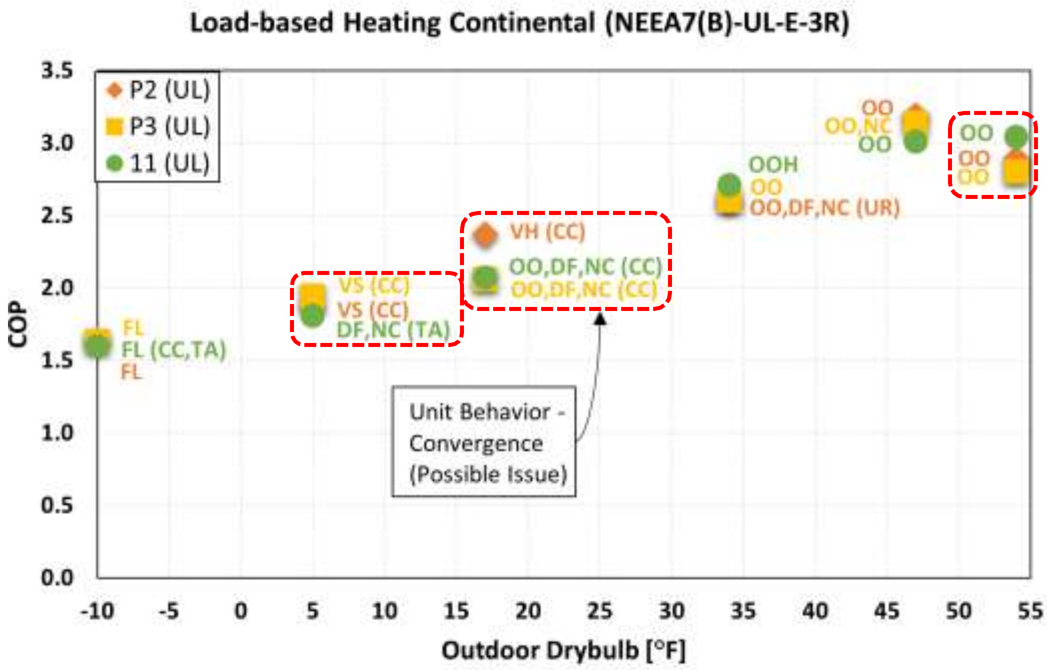


Figure 80. NEEA7(B) (UL) heating continental test intervals COP for different tests with unit behavior during convergence

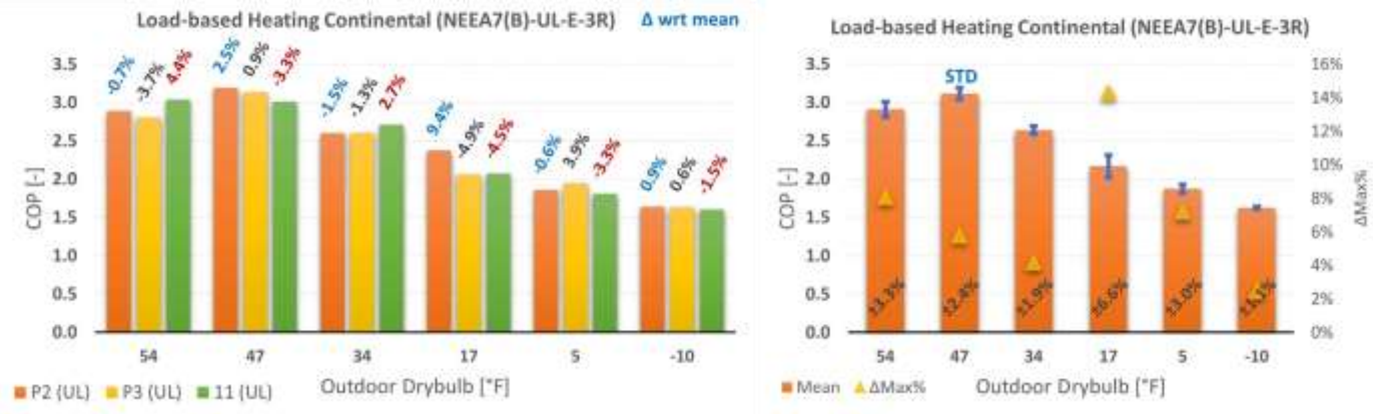


Figure 81. NEEA7(B) (UL) heating continental test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

In 17°F test interval (HC), similar unit dynamic behavior with on/off cycles and defrost was observed in three tests as shown in Figure 82, Figure 83, and Figure 84 for tests P2, P3, and #11, respectively. However, in test P2 the convergence was found at the beginning of the test before the unit cycled off (Figure 82) and the convergence period did not consider the on/off and defrost cycles in overall performance as in test P3 (Figure 83) and test #11 (Figure 84). This resulted in a relatively higher COP in test P2 compared to other tests. So, the main cause for this large difference in test interval HC is due to convergence criteria, which should be updated to account for the cases like this to capture representative test unit performance. Also, higher airflow was observed in test #11, which does not seem to have a large impact as comparable COP was observed in tests P3 and #11.

At 54°F test interval, a similar on/off cycling unit response was observed, however, in test #11 around 5.1% higher COP was measured compared to test P2 and around 8.1% higher compared to test P3. This difference could be attributed to variation in test unit dynamic response and variability after test unit re-installation in addition to measurement and dynamic testing uncertainties. For test interval at 5°F ambient temperature test unit utilized defrost, however, convergence was found in steady mode 40 minutes window for tests P2 and P3, thus not accounted for the defrost in test interval performance unlike converged performance in test #11. This resulted in a relatively lower COP in test #11 compared to tests P2 and P3. A similar issue was observed in the NEEA4 test results. Convergence criteria should be updated to account for defrost for cases like this. In addition, in test #11, only one complete defrost cycle was captured in 6 hours, and the test was terminated before the second cycle finishes as per the current time limit of 6 hours in EXP07 for a heating test. This should be investigated whether 6 hours' time limit in the current testing approach is appropriate or not. Also, interestingly in the lowest ambient temperature full-load test interval, the converged period did not include defrost in all three cases as convergence was found in a steady-state 40 min window.

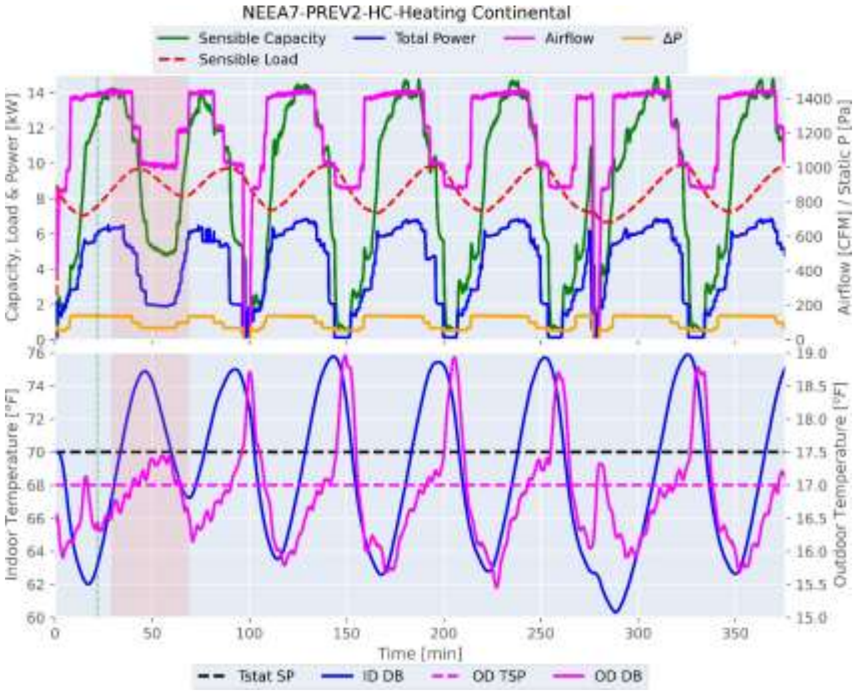


Figure 82. NEEA7-P2 (UL) performance and temperature variations for heating continental test interval – HC (17°F)

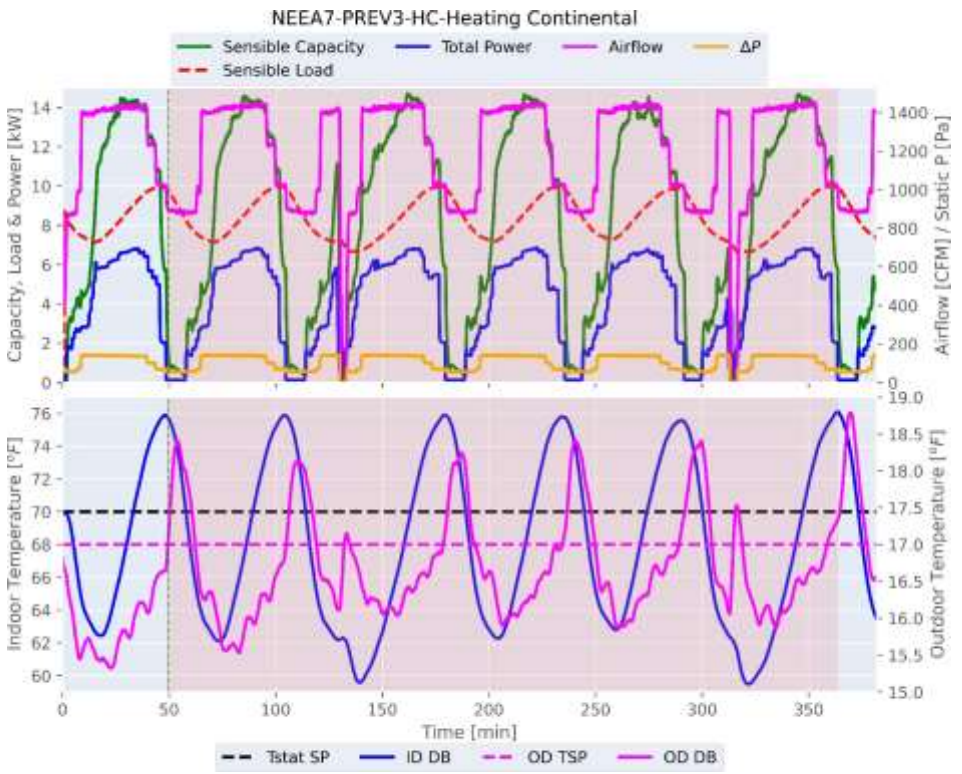


Figure 83. NEEA7-P3 (UL) performance and temperature variations for heating continental test interval - HC (17°F)

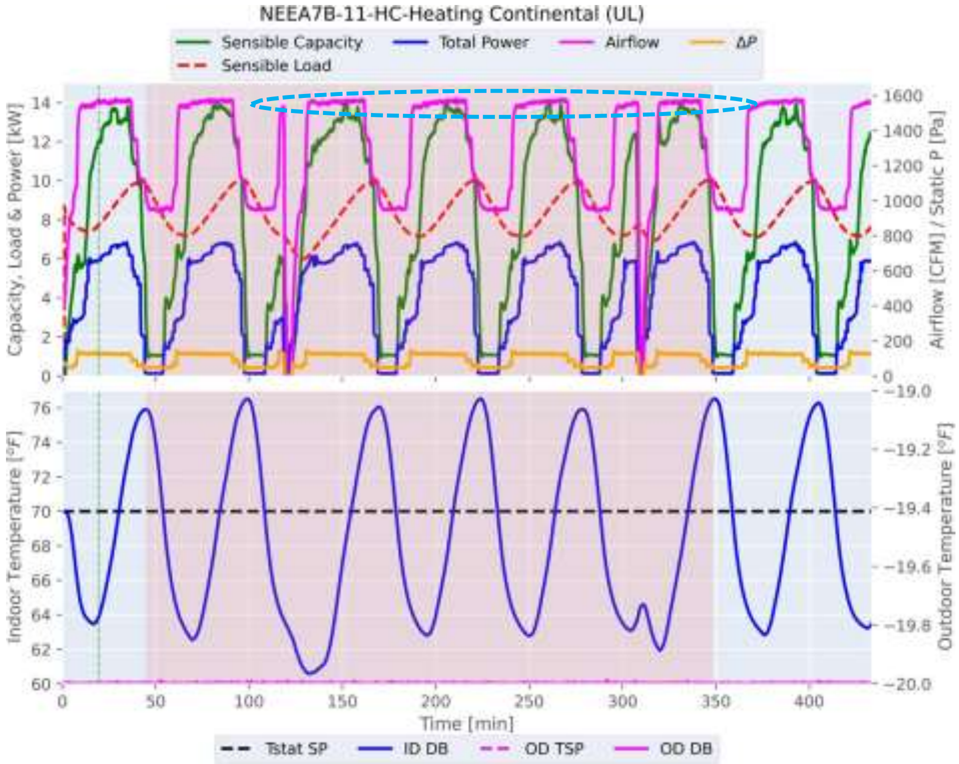


Figure 84. NEEA7B-11 (UL) performance and temperature variations for heating continental test interval - HC (17°F)

Figure 85 and Figure 86 show the repeatability comparison of heating marine test intervals performance. At the same test conditions, a similar unit operation mode was captured during the convergence period except at the 17°F test interval, where a larger variation in performance was observed similar to heating continental test results. The maximum difference in COP varies from 4.1% to 10.1% and STD ±1.7% to ±4.5%, slightly better than continental test intervals. In heating marine test intervals, possible issues related to the convergence criteria during a mix of on/off and defrost cycles were observed similar to as described above for continental conditions. At 17°F ambient temperature, the unit cycled on/off with intermittent defrosts in tests P2 and P3; whereas, in test #11, the unit operated in variable speed hybrid mode in between defrosts rather than on/off cycles. Also, the airflow in test 11 was higher compared to the other two tests, which could be due to a possible test setup issue or just an inconsistent unit dynamic response. In addition to the difference in unit dynamic response, the difference in the convergence period also contributed to an overall larger difference in COP. In test P2, convergence was found in 40 minutes steady operation period and in test P3 during two successive on/off cycles without defrost in the convergence window. Whereas in test #11, performance did not convergence and the entire 6-hour test duration data was taken in average performance, which is not right as only complete on/off or defrost cycles should have been taken in convergence.

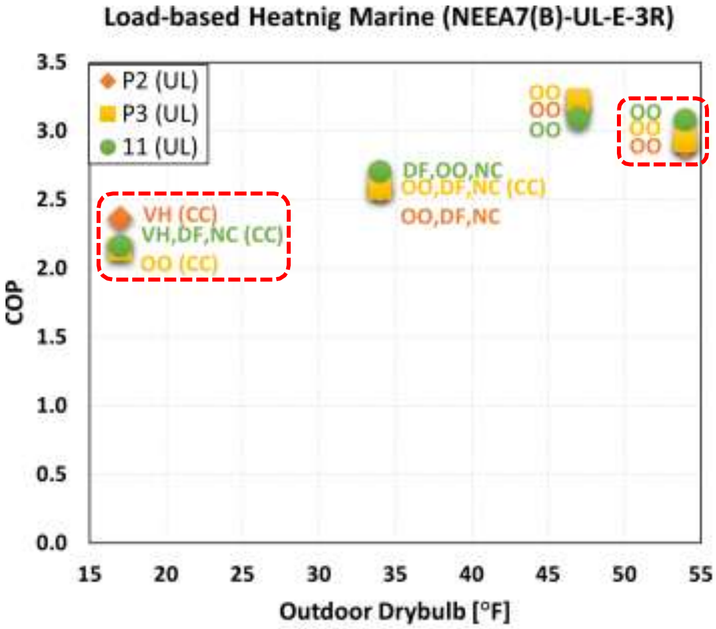


Figure 85. NEEA7(B) (UL) heating marine test intervals COP for different tests with unit behavior during convergence

Figure 87 and Figure 88 show comparisons of NEEA7(B) estimated heating SCOP for different climate zones based on three sets of tests at UL in dataset NEEA7(B)-UL-E-3R. For the "Marine" climate zone, which utilizes marine test interval results in SCOP estimation, good repeatability was observed with a maximum difference in SCOP of 0.8% between the three tests. For climate zones that utilize continental interval test results, the maximum difference in SCOP varies from 2% to 4.7% and STD from ±0.8% to ±2.1%, overall showing good repeatability.

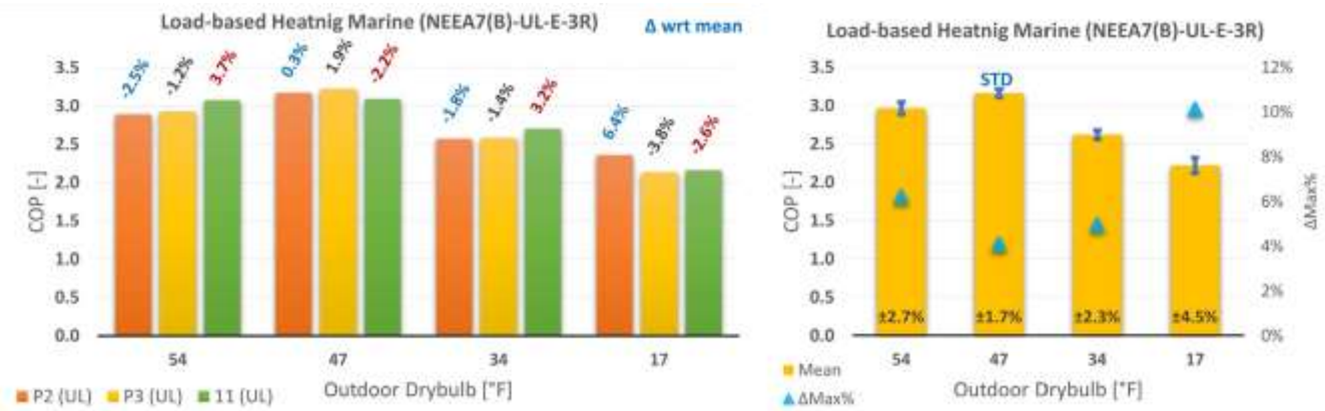


Figure 86. NEEA7(B) (UL) heating marine test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

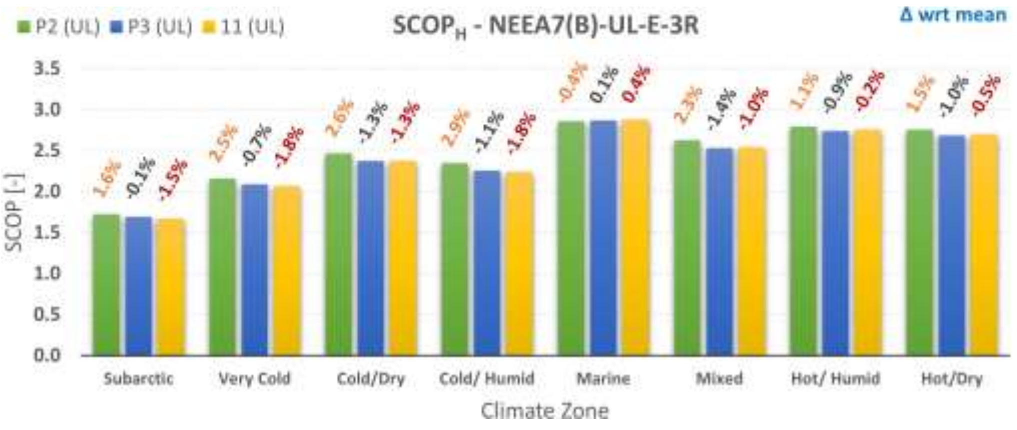


Figure 87. NEEA7(B) (UL) heating SCOP comparison in different climate zones for different tests

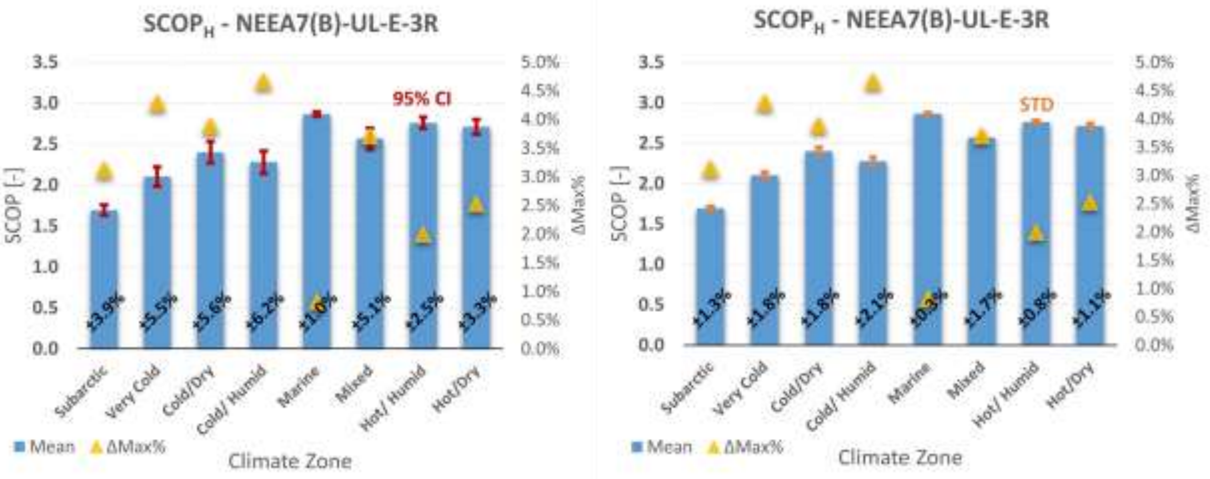


Figure 88. NEEA7(B) (UL) heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

Table 16 shows the NEEA7(B) overall SCOP repeatability results summary based on three sets of test results at UL in dataset NEEA7(B)-UL-E-3R. Relatively better repeatability was observed in heating mode compared to cooling mode. The differences are thought to be mainly due to inconsistencies in unit control behavior and the need to improve convergence criteria and their implementation. Table 17 shows the NEEA7 repeatability summary results for dataset NEEA7-UL-E-2W with only considering two back-to-back sets of tests (P2 and P3) without test #11 which was performed after test unit re-installation. Without considering test #11 in the assessment improves repeatability comparably for cooling, however, no significant change in heating results.

Table 16. NEEA7(B)-UL-E-3R dataset SCOP repeatability summary results

Metric	No. of Tests	∆SCOP (Max-Min) [%]			SCOP 95% CI [%]			SCOP STD [%]		
		Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
$SCOP_C$	3	5.1%	6.8%	5.9%	6.4%	8.5%	7.4%	2.1%	2.8%	2.4%
$SCOP_H$	3	0.8%	4.7%	3.1%	1.0%	6.2%	4.1%	0.3%	2.1%	1.4%

Table 17. NEEA7-UL-E-2W dataset SCOP repeatability summary results without test #11

Metric	No. of Tests	∆SCOP (Max-Min) [%]			SCOP 95% CI [%]			SCOP STD [%]		
		Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
$SCOP_C$	2	1.5%	3.3%	2.5%	9.3%	20.9%	15.6%	0.7%	1.6%	1.2%
$SCOP_H$	2	0.5%	3.9%	2.7%	3.2%	24.9%	17.0%	0.2%	2.0%	1.3%

4.5 NEEA7B (PG&E)

In this section, EXP07 repeatability assessment analysis results are presented for NEEA7B, a 3-ton split-type variable-speed heat pump system, based on two sets of test results at the PG&E lab with dataset NEEA7B-PGE-E-2W.

4.5.1 Cooling Mode

Figure 89 and Figure 90 show the comparison of cooling dry coil and humid coil test intervals COP, respectively, with unit operation mode during the convergence period and possible issues with test interval for two repeated tests of the NEEA7B unit at PG&E in dataset NEEA7B-PGE-E-2W. It should be noted that at PG&E, a different unit but the same model as used in tests P2 and P3 at UL was tested due to some issues with the original unit shipped from UL. At the same test conditions, a similar unit operation mode was captured during the convergence period in repeated tests for cooling mode performance measurement. Figure 91 shows the mean COP of repeated tests for dry and humid coil test intervals with STD and the absolute COP percentage difference for a quantitative measure of repeatability. For dry coil test intervals, the absolute difference in COP between two tests at the same test conditions varies from 0% to 6.3% and STD from ±0% to ±3.1%. Slightly better repeatability was observed in humid coil test intervals, with COP difference varying from 0.9% to 1.6% and STD from ±0.4% to ±0.8%. In general, the test unit performed better in test #4 compared to test #2 across most of the test intervals. Overall test unit showed good repeatability in cooling mode performance at PG&E, with the largest differences observed in 77°F and 86°F outdoor temperature dry coil test intervals where the unit was cycling on/off. At these conditions, the unit showed similar dynamic behavior in repeated tests, however, the difference could be because of some variation in unit dynamic response during compressor ramp up/down during cycle *on* period due to some variability in unit control response and thermostat sensing dynamics.

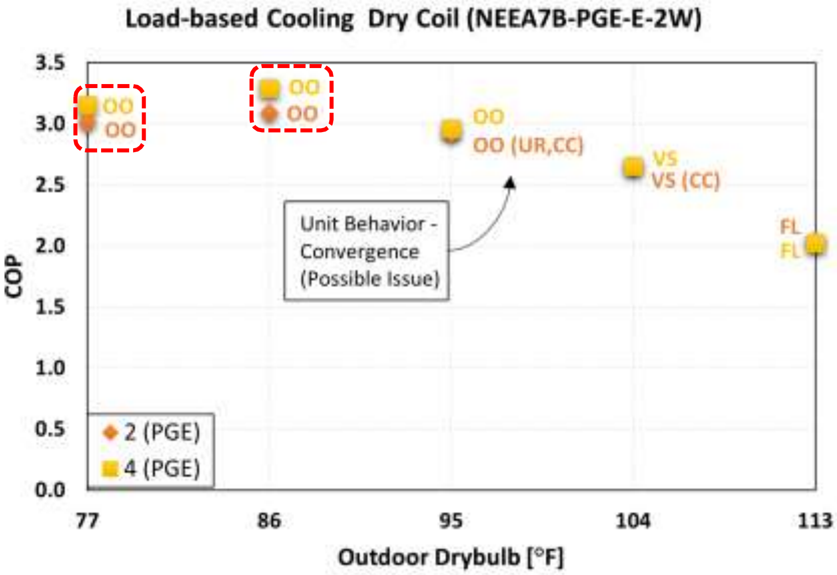


Figure 89. NEEA7B (PG&E) cooling dry coil test intervals COP for different tests with unit behavior during convergence

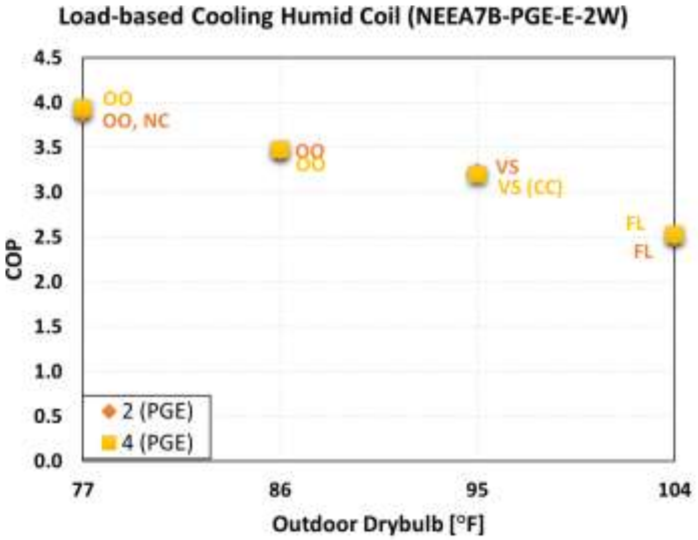


Figure 90. NEEA7B (PG&E) cooling humid coil test intervals COP for different tests with unit behavior during convergence

Further, some possible issues related to convergence criteria and inconsistent unit dynamic response were observed. For example, in test #2 dry coil test interval at 95°F outdoor temperature, as shown in Figure 92, the test unit was mainly operating in variable speed hybrid mode in the initial 12 hours with one cycle off around the 4th hour. Then around the 13th hour, the unit started cycling on/off in a somewhat regular pattern which is similar to what was observed in test #4 at the same test conditions from the beginning of the test as shown in Figure 93. Considering regular on/off cycles in convergence during test #2 interval and ignoring the initial 12 hours resulted in comparable COP between two tests under the same conditions. However, implementing the convergence criteria strictly as per EXP07 would have resulted in no convergence after a 4-hour test duration in test #2 and we would have taken the average performance those 4 hours as the test interval performance. That could have resulted in quite different performances between two repeated tests. This example showcases the possible issues related to the unit's inconsistent dynamic response that could affect repeatability or reproducibility under the same test conditions. The possible convergence criteria issue indicated for 104°F dry coil and 95°F humid coil test interval is associated with convergence criteria not capturing representative performance when test unit showing variable speed hybrid mode i.e., compressor ramping up and down in a regular pattern without cycling off. This is similar to what was observed in some test intervals at UL and discussed above.

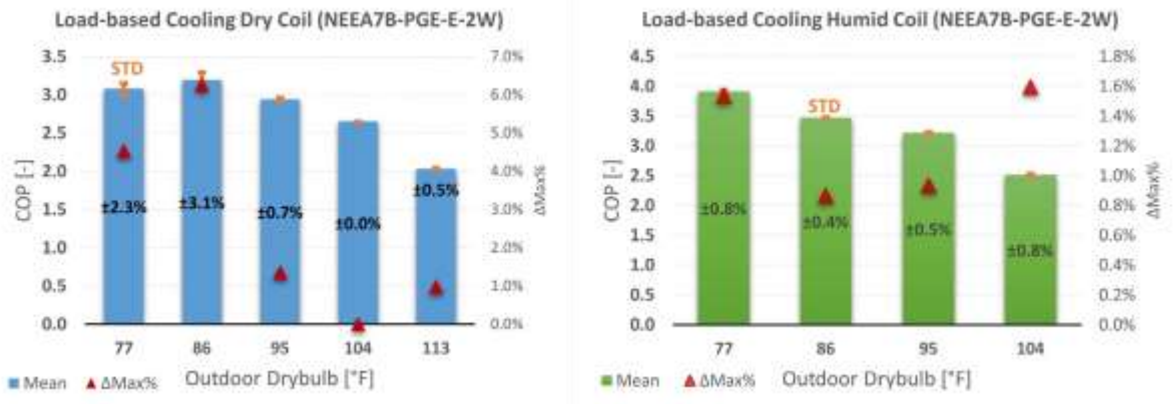


Figure 91. NEEA7B (PG&E) cooling dry coil and humid coil test intervals mean COP with STD and maximum difference among repeated tests

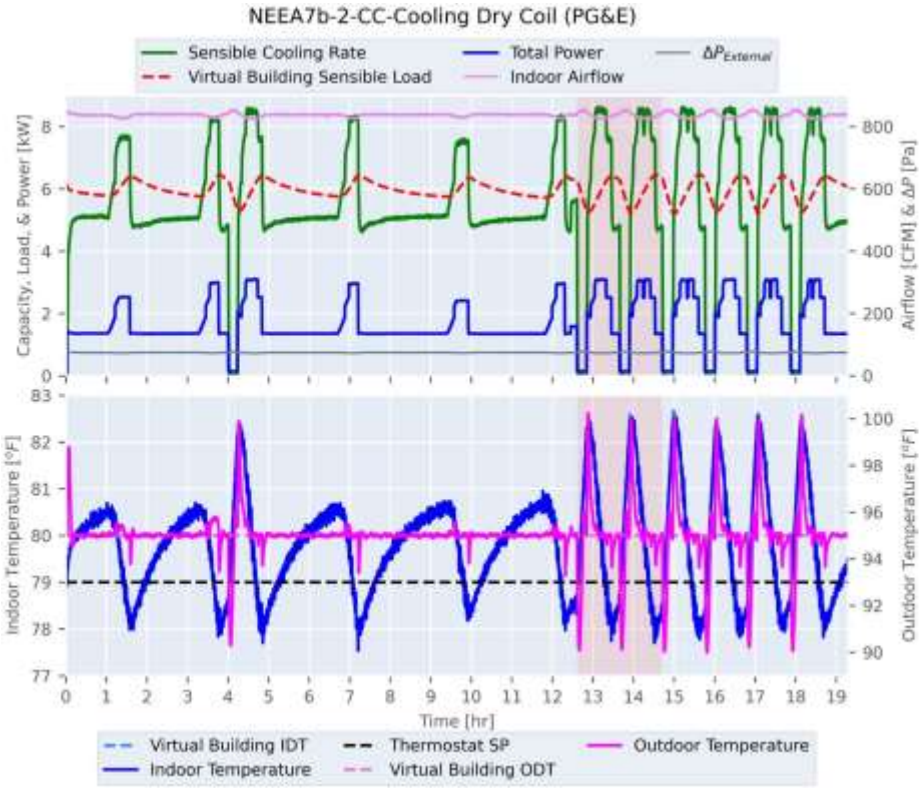


Figure 92. NEEA7B-2 (PG&E) performance and temperature variations for cooling dry coil test interval CC

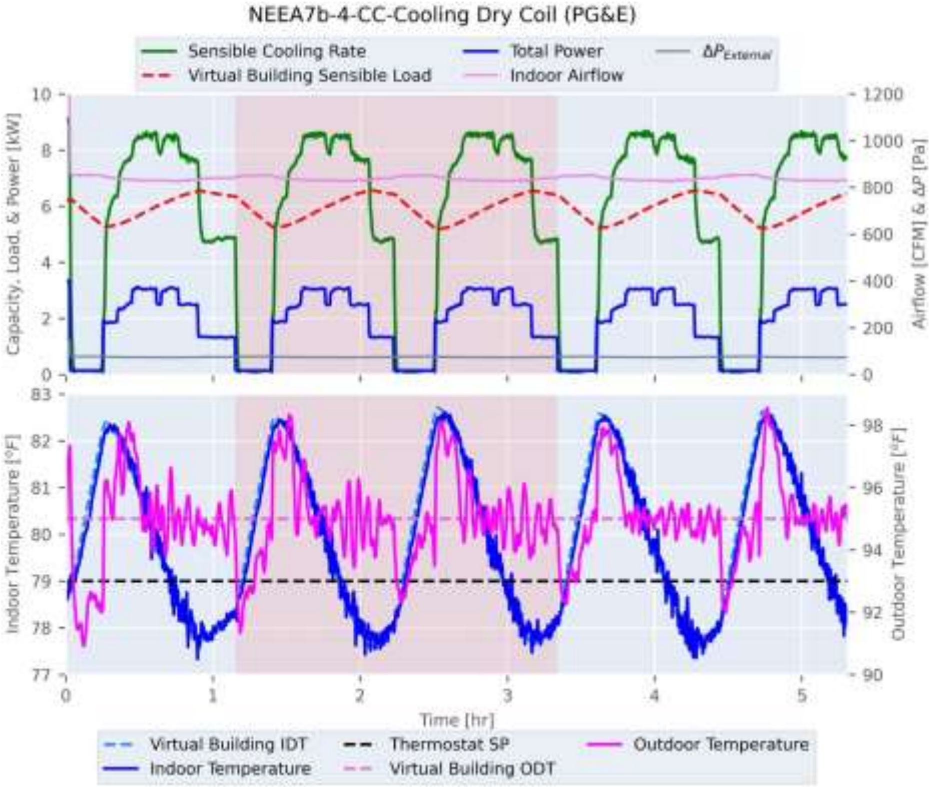


Figure 93. NEEA7B-4 (PG&E) performance and temperature variations for cooling dry coil test interval CC

Figure 94 and Figure 95 show the comparison of NEEA7B estimated cooling SCOP for different climate zones based on measured performance in two sets of different tests at PG&E for dataset NEEA7B-PGE-E-2W. Overall good repeatability is observed in cooling SCOP. For climate zones that utilize cooling dry coil test results in SCOP estimation (3 out of 7), the maximum difference varies from 2.9% to 3.8% and STD from ±1.5% to ±1.9%. For climate zones utilizing humid coil test results, a relatively smaller variation was observed with the maximum SCOP difference around 0.8% and STD around ±0.4%. Even with smaller differences, 95% confidence intervals are estimated somewhat large, because only two samples (sets of tests) were used here.

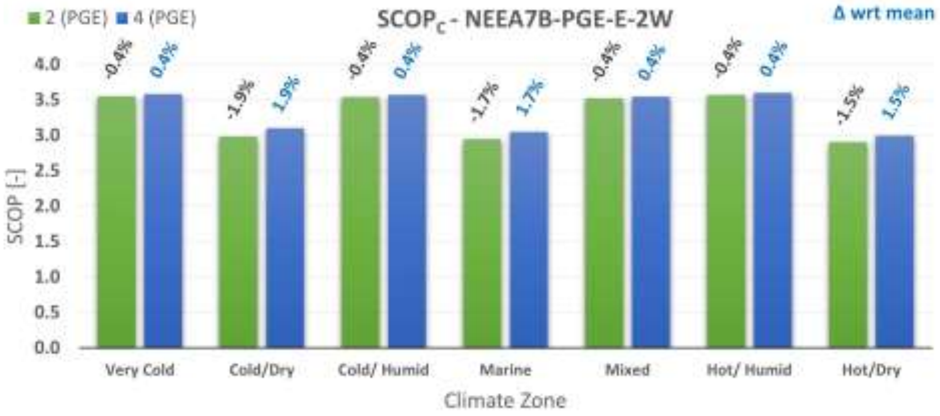


Figure 94. NEEA7B (PG&E) cooling SCOP comparison in different climate zones for different tests

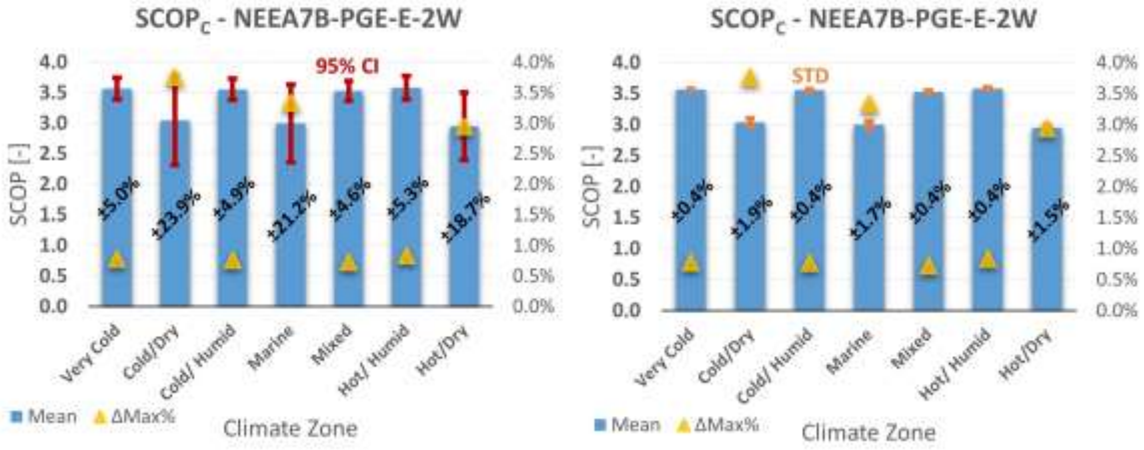


Figure 95. NEEA7B (PG&E) cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

4.5.2 Heating Mode

Figure 96 and Figure 97 show the comparison of NEEA7B performance in heating continental and marine test intervals, respectively, for 2 sets of repeated tests with dataset NEEA7B-PGE-E-2W. One thing to note is that at PG&E, test unit performance was not measured at the lowest outdoor temperature (-10°F) test interval similar to NEEA 4 case due to test facility limitations. During the convergence period, similar unit operation modes were observed in the same test intervals of repeated tests, except for the 34°F outdoor temperature continental test interval where one defrost was captured in between on/off cycles for test #2 and only on/off cycles in test #4. Also, at 34°F continental and marine test intervals, convergence was achieved in one test but not in another. During full-load tests in 17°F and 5°F ambient temperature continental test intervals, even though the unit was going into defrost, the convergence was found in steady operation in between defrosts and thus average performance did not account for defrost which is related to possible convergence criteria issue as discussed before. Also in these test intervals, a possible testing approach issue was noted related to not running the test long enough to capture multiple complete defrost cycles.

Nevertheless, as shown in Figure 98, the test unit showed overall good repeatability in heating tests with an absolute difference in COP varying from 0.1% to 2.5% and STD from ±0.1% to ±1.2% for continental intervals, and for marine intervals, relatively larger differences observed with an absolute difference in COP varying from 0.9% to 5.7% and STD from ±0.5% to ±2.8%. The largest difference in COP between repeated tests was noted in full-load tests at 17°F marine test intervals. This seems mainly due to the difference in test unit response to utilize defrost as shown in Figure 99 and Figure 100 for tests #2 and #3, respectively. In test #2 unit showed consistent defrost cycles with a similar period; however, in test #3 some short defrost cycles were also observed in between longer ones. This resulted in a measured COP of around 5.7% lower in test #3 compared to test #2. The other interesting thing to notice is relatively large fluctuations in outdoor temperatures during defrosts and in indoor temperature during these full-load tests which will affect repeatability and reproducibility, and also has implications to test condition and operating tolerances.

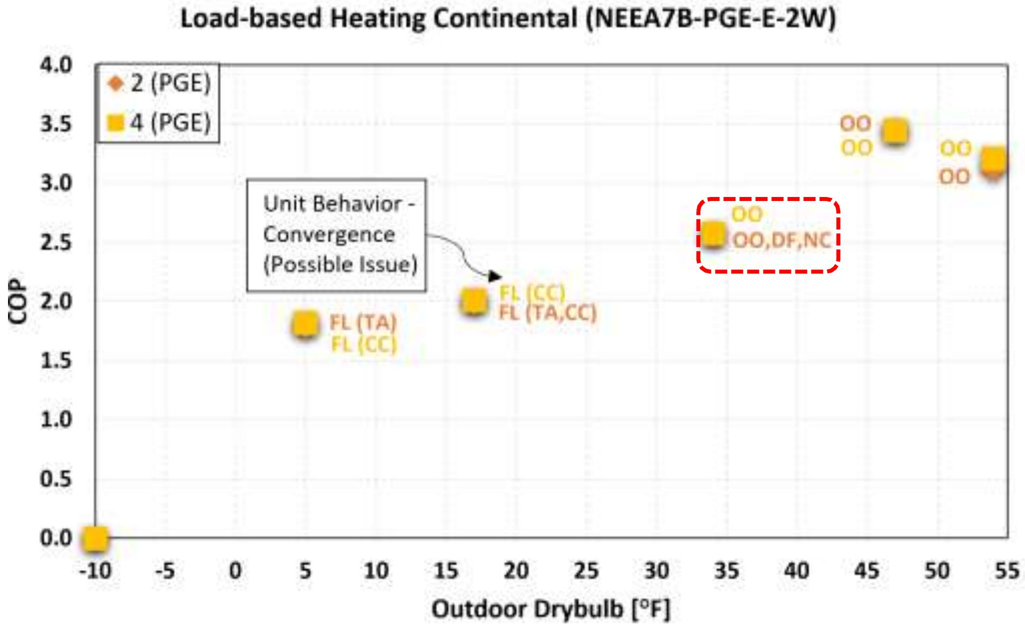


Figure 96. NEEA7B (PG&E) heating continental test intervals COP for different tests with unit behavior during convergence

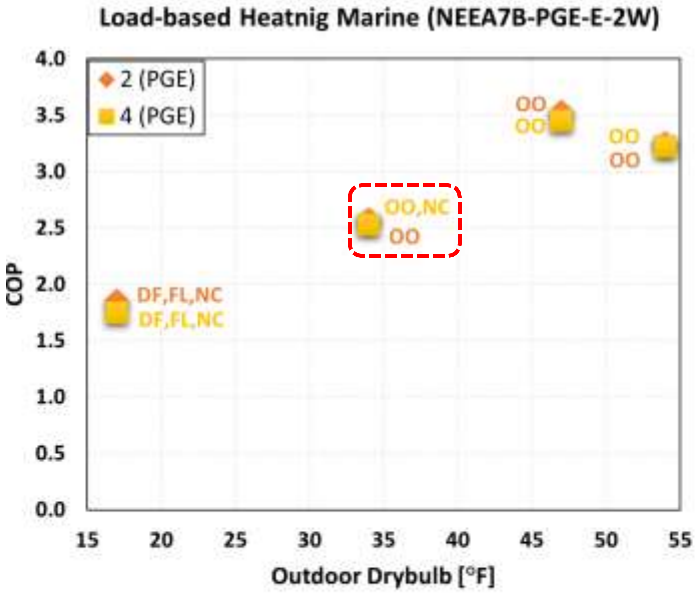


Figure 97. NEEA7B (PG&E) heating marine test intervals COP for different tests with unit behavior during convergence

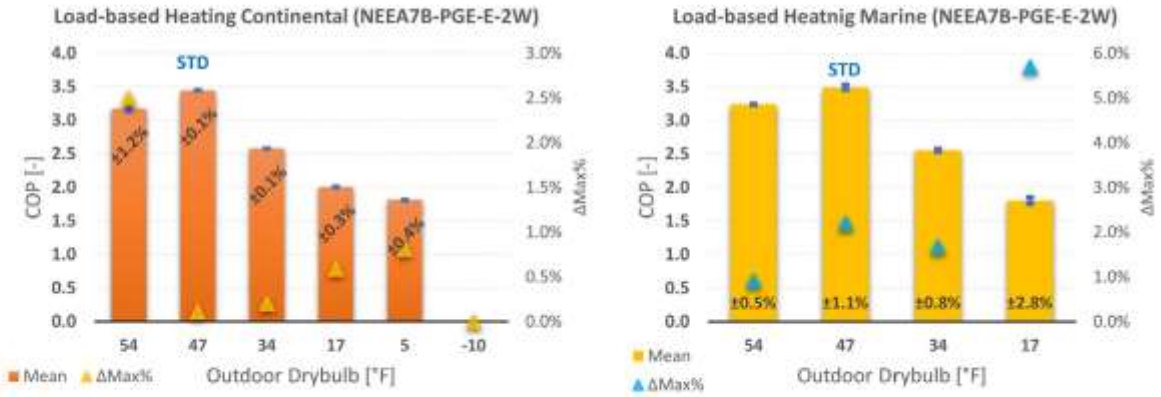


Figure 98. NEEA7B (PG&E) heating continental and marine test intervals mean COP with STD and maximum difference among repeated tests

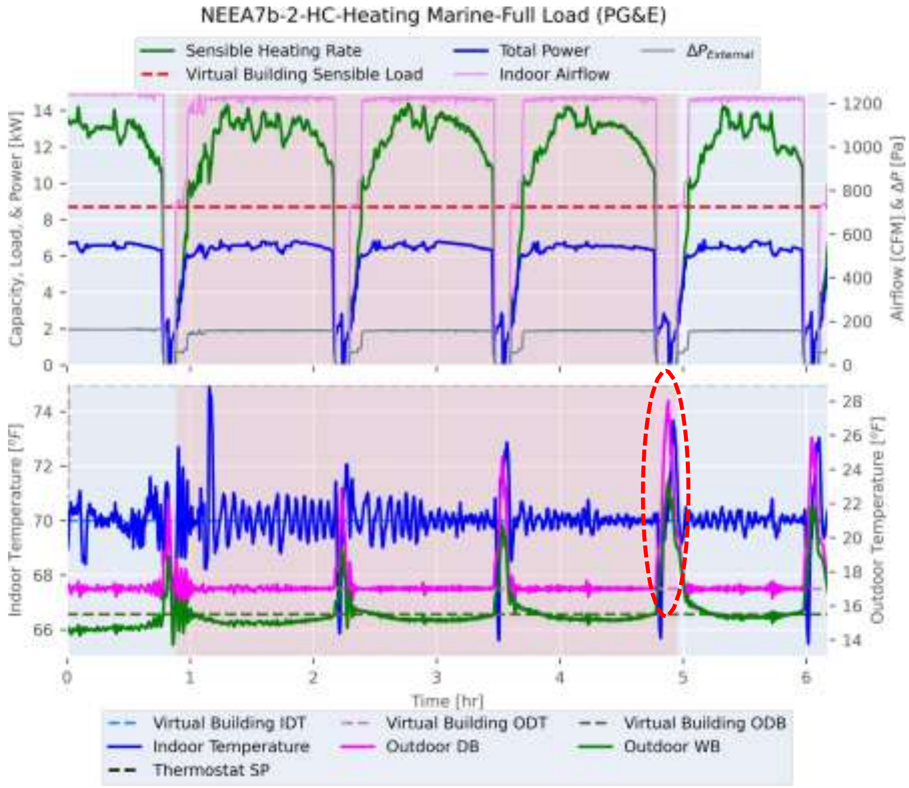


Figure 99. NEEA7B-2 (PG&E) performance and temperature variations for heating marine test interval HC (17℉)

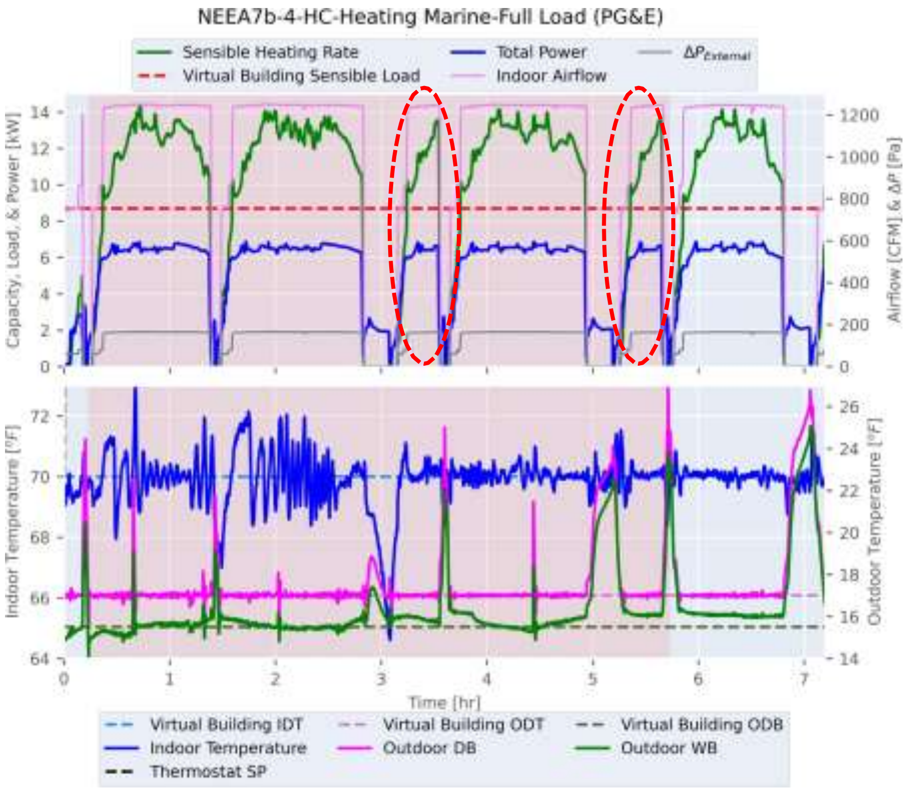


Figure 100. NEEA7B-4 (PG&E) performance and temperature variations for heating marine test interval HC (17°F)

Figure 101 and Figure 102 show NEEA7B estimated heating SCOP comparison for different climate zones along with the maximum differences and average SCOP with a 95% CI and STD based on two sets of tests at PG&E with dataset NEEA7B-PGE-E-2W. Overall good repeatability was noted with differences in heating SCOP varying from 0.2% to 2.1% and STD from ±0.1% to ±1%, with the largest difference observed in climate zone which utilizes marine interval test results.

Table 18 shows the overall SCOP repeatability results summary with the range and mean of three statistical parameters for NEEA7B at PG&E. Overall test unit showed good repeatability with comparatively better in heating mode compared to cooling mode test results.

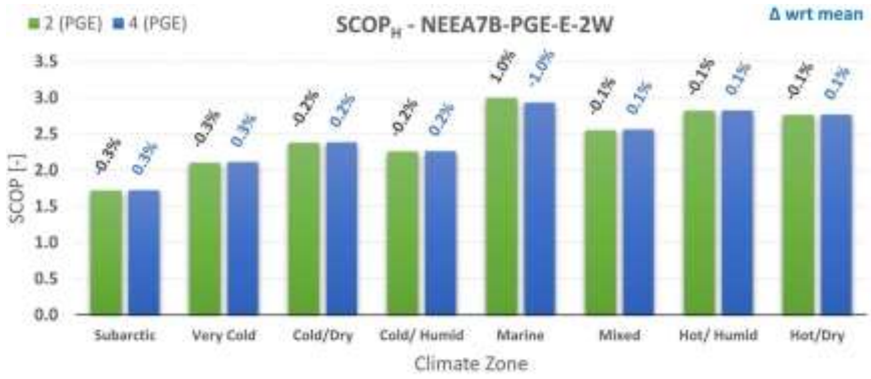


Figure 101. NEEA7B (PG&E) heating SCOP comparison in different climate zones for different tests

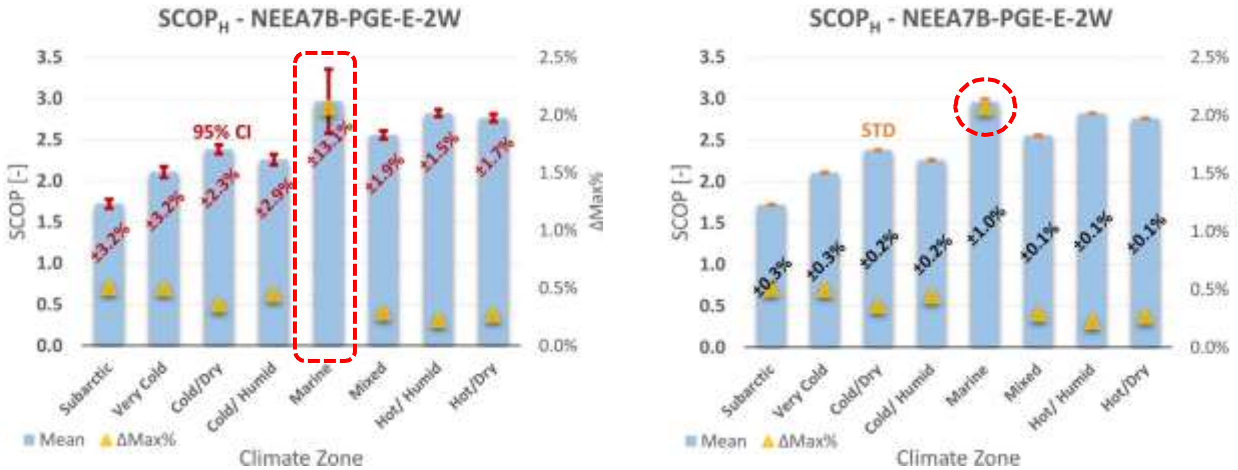


Figure 102. NEEA7B (PG&E) heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

Table 18. NEEA7B-PGE-E-2W dataset SCOP repeatability summary results

Metric	**No. of Tests**	**ΔSCOP (Max-Min) [%]**			**SCOP 95% CI [%]**			**SCOP STD [%]**		
		Min	***Max***	***Mean***	***Min***	***Max***	***Mean***	***Min***	***Max***	***Mean***
SCOP$_C$	2	0.7%	3.8%	1.9%	4.6%	23.9%	12.0%	0.4%	1.9%	0.9%
SCOP$_H$	2	0.2%	2.1%	0.6%	1.5%	13.1%	3.7%	0.1%	1.0%	0.3%

4.6 NRCan8 (UL)

In this section, EXP07 repeatability assessment analysis results are presented for NRCan8, a 1.5-ton mini-split ductless variable-speed heat pump system, based on two sets of test results at the UL lab with dataset NRCan8-UL-E-2W.

4.6.1 Cooling Mode

Figure 103 and Figure 104 show the performance and unit behavior comparison of cooling dry coil and humid coil test intervals, respectively, with possible issues with test intervals for two back-to-back repeated tests of NRCan8 unit at UL for dataset NRCan8-UL-E-2W. During the convergence period, similar unit operation modes were observed at the same test conditions. Figure 105 shows the mean COP of repeated tests for dry and humid coil test intervals with STD and the absolute COP percentage difference on the right vertical axis. For dry coil test intervals, the absolute difference in COP between two tests at the same test conditions varies from 0.4% to 6% and STD from ±0.2% to ±3%. Somewhat better repeatability was observed in humid coil test intervals, with COP difference varying from 0.8% to 1.3% and STD from ±0.4% to ±0.6%. Overall test unit showed good repeatability in cooling mode performance. The largest COP differences were observed in 77°F and 86°F outdoor temperature dry coil test intervals of 6% and 5.4%, respectively, where the unit performed better in test #1 compared to test #2. At these conditions, the unit showed similar dynamic behavior in repeated tests, however, the difference could be because of some variation in unit dynamic response during compressor ramp up/down due to some variability in unit control response and thermostat sensing dynamics. The other possible cause of the difference is some variation in the period where convergence was found which is again related to variation in test unit dynamic response with time at the same test conditions.

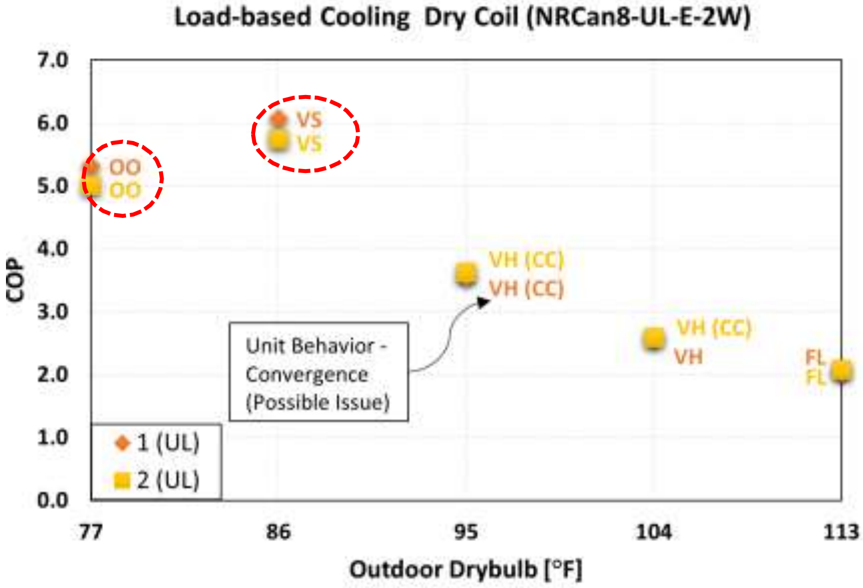


Figure 103. NRCan8 (UL) cooling dry coil test intervals COP for different tests with unit behavior during convergence

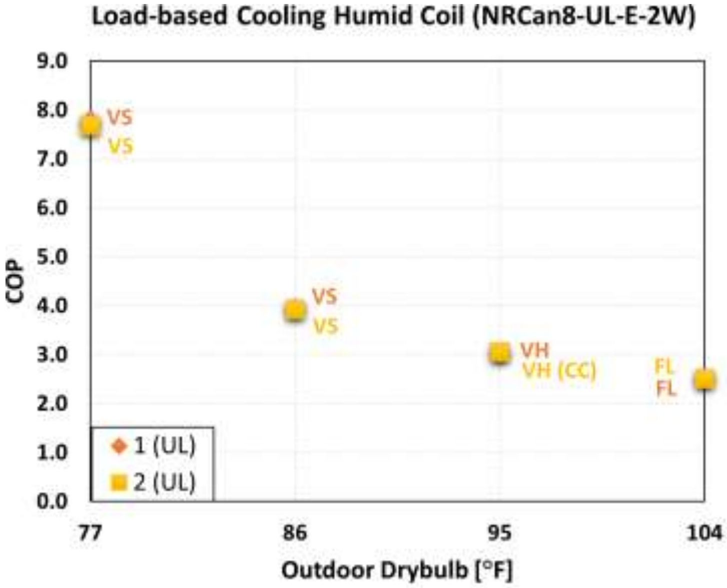


Figure 104. NRCan8 (UL) cooling humid coil test intervals COP for different tests with unit behavior during convergence

Further, a possible issue related to convergence criteria was identified for some test intervals where the unit was operating in variable speed hybrid mode. For example, in test #2 dry coil test interval at 104°F outdoor temperature (CB) as shown in Figure 106, the test unit was operating in variable speed hybrid mode and convergence was found in a 40 min window at the beginning of the test.

However, the convergence criterion is not capturing representative performance for this test interval, which would be capturing the complete compressor ramp up and down cycle in average performance. This is similar to what was observed in some test intervals for other test units also and discussed previously.

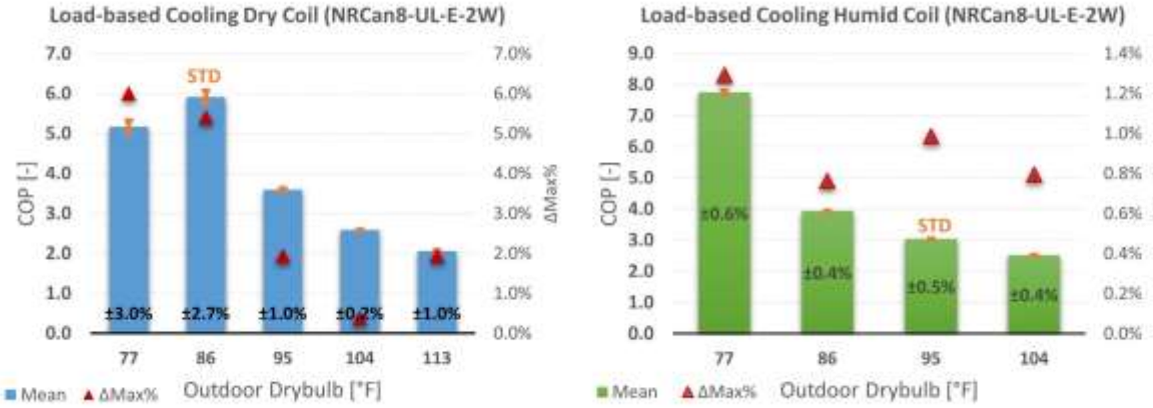


Figure 105. NRCan8 (UL) cooling dry coil and humid coil test intervals mean COP with STD and maximum difference among repeated tests

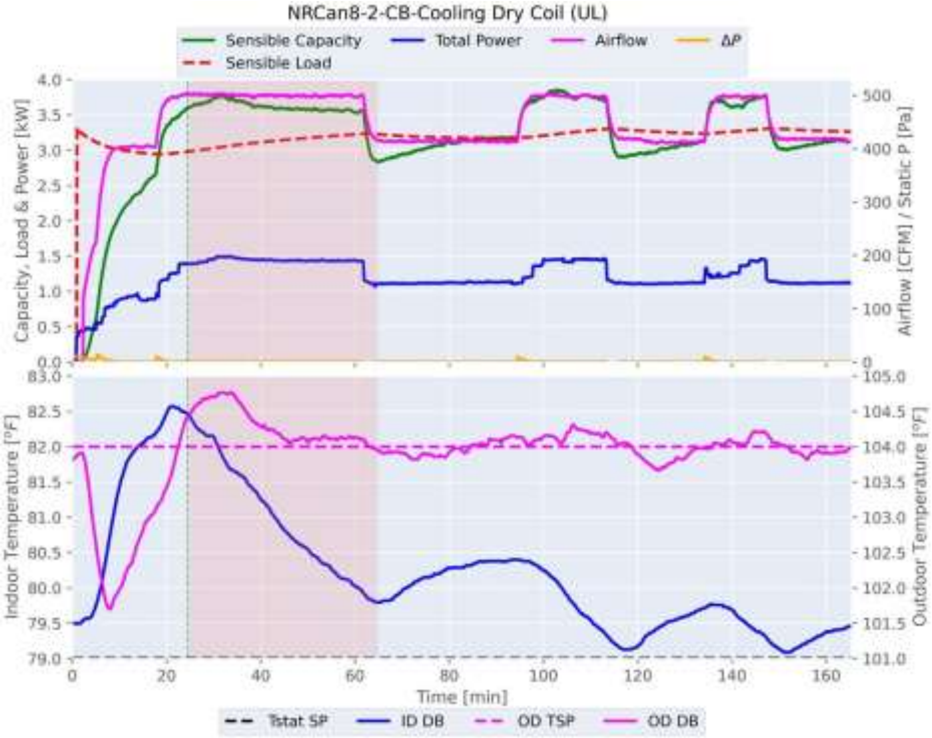


Figure 106. NRCan8-2 (UL) performance and temperature variations for cooling dry coil test interval CB

Figure 107 and Figure 108 show the comparison of NRCan8 estimated cooling SCOP for different climate zones between two sets of tests for dataset NRCan8-UL-E-2W. Overall good repeatability was observed in cooling SCOP, with maximum differences varying from 0.6% to 1.7% and STD from ±0.3% to ±0.8% across different climate zones.

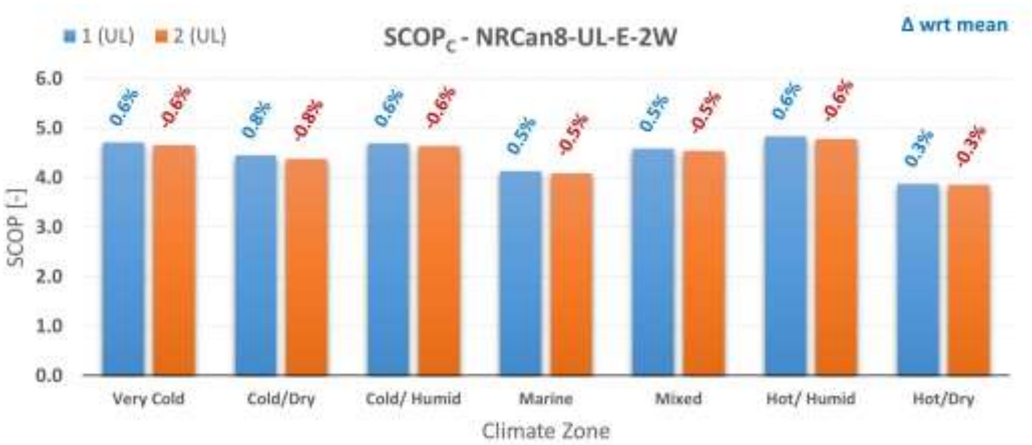


Figure 107. NRCan8 (UL) cooling SCOP comparison in different climate zones for different tests

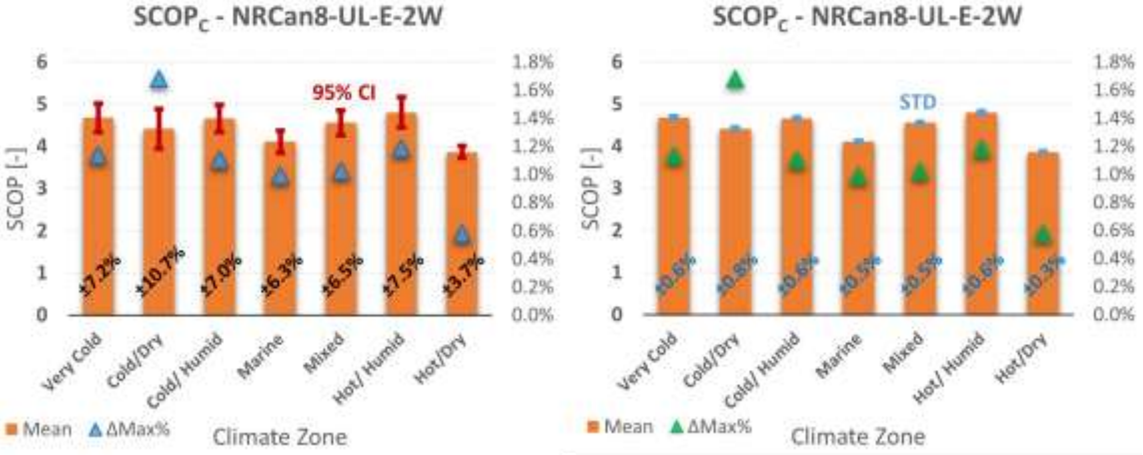


Figure 108. NRCan8 (UL) cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

4.6.2 Heating Mode

Figure 109 and Figure 110 show NRCan8 performance comparison in heating continental and marine test intervals, respectively, for 2 sets of repeated tests. Similar unit operation modes were observed during convergence at the same test conditions in repeated tests. One interesting thing to note is that in none of the heating test intervals, test unit defrost performance and behavior was captured, which is due to some possible issues related to convergence criteria and testing approach in test intervals where the unit utilized defrost. For example, in the 34°F ambient temperature heating continental test interval, for test #2 as shown in Figure 112, the test unit was going into defrost, however, the convergence was found between two defrosts in a 40 min window. Thus, average performance did not account for defrosts which is not representative of this test interval performance. Convergence criteria issue at lower ambient temperature in marine and continental test intervals is similar to this and the testing approach issue in these test intervals is related to not running a test long enough to capture multiple complete defrost cycles. Further, a possible issue with inconsistent unit dynamic response was observed in 54°F marine test intervals, where the unit showed short on/off cycling behavior as shown in Figure 113. However, in same ambient temperature continental test interval unit showed a regular on/off cycling behavior as shown in

Figure 114. The only difference between the two test intervals is outdoor humidity which should not affect test unit cycling behavior in this case without defrost. So, this inconsistency in cycling behavior was possibly due to variation in test unit control response and thermostat sensing dynamics, which could affect load-based tests' repeatability and reproducibility.

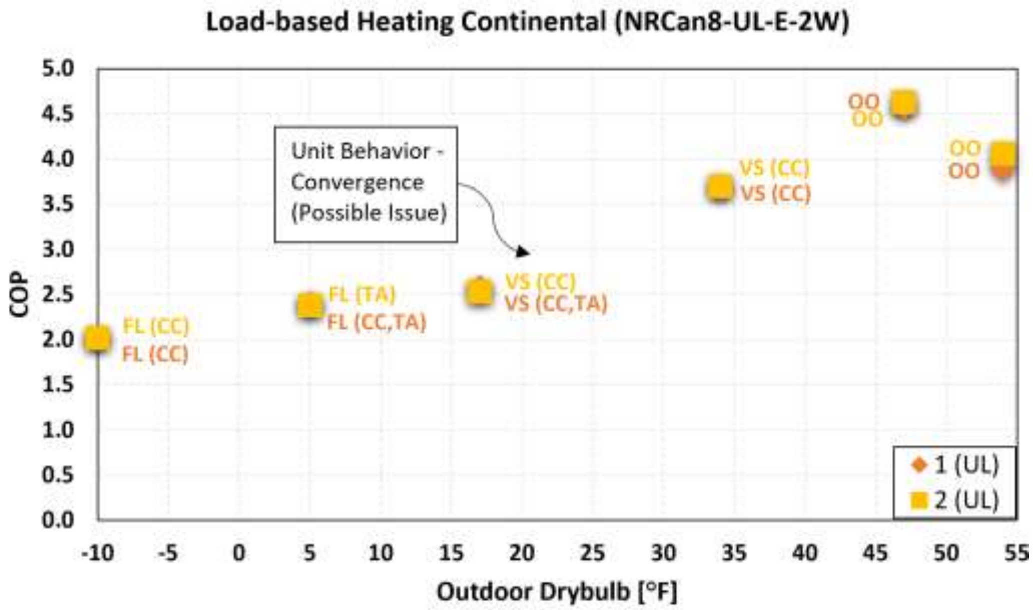


Figure 109. NRCan8 (UL) heating continental test intervals COP for different tests with unit behavior during convergence

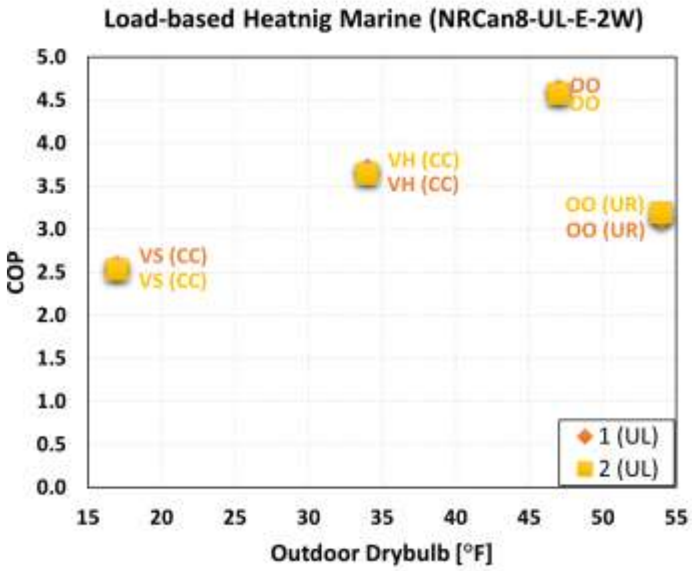


Figure 110. NRCan8 (UL) heating marine test intervals COP for different tests with unit behavior during convergence

Figure 111 shows the mean COP of repeated tests for continental and marine test intervals with STD and the absolute COP percentage difference. Overall, the test unit showed good repeatability in heating tests with an absolute difference in COP varying from 0% to 4.0% and STD from ±0%

to ±2% for continental intervals, and for marine intervals, the absolute difference in COP varies from 0.8% to 1.6% and STD from ±0.4% to ±0.8%. The largest difference of 4% in COP between repeated tests was noted at 54°F continental test interval, where the unit was cycling on/off. This is due to some variation in test unit dynamic response and convergence period. In test #1 convergence was found in the 2nd and 3rd cycles (Figure 114), but in test #2, the 7th and 8th cycles converged and before that, the unit seemed to go in timed defrost even at this high ambient temperature as shown in Figure 115.

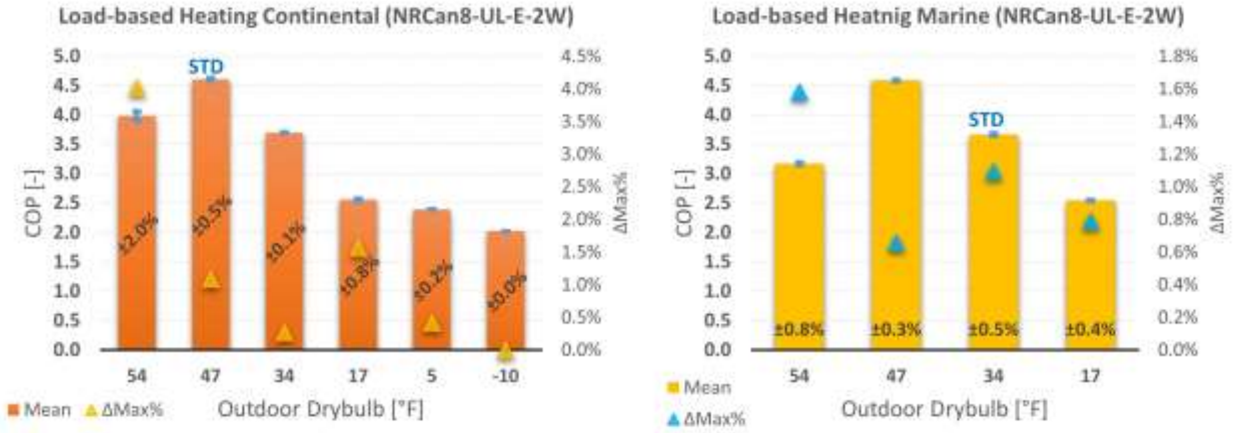


Figure 111. NRCan8 (UL) heating continental and marine test intervals mean COP with STD and maximum difference among repeated tests

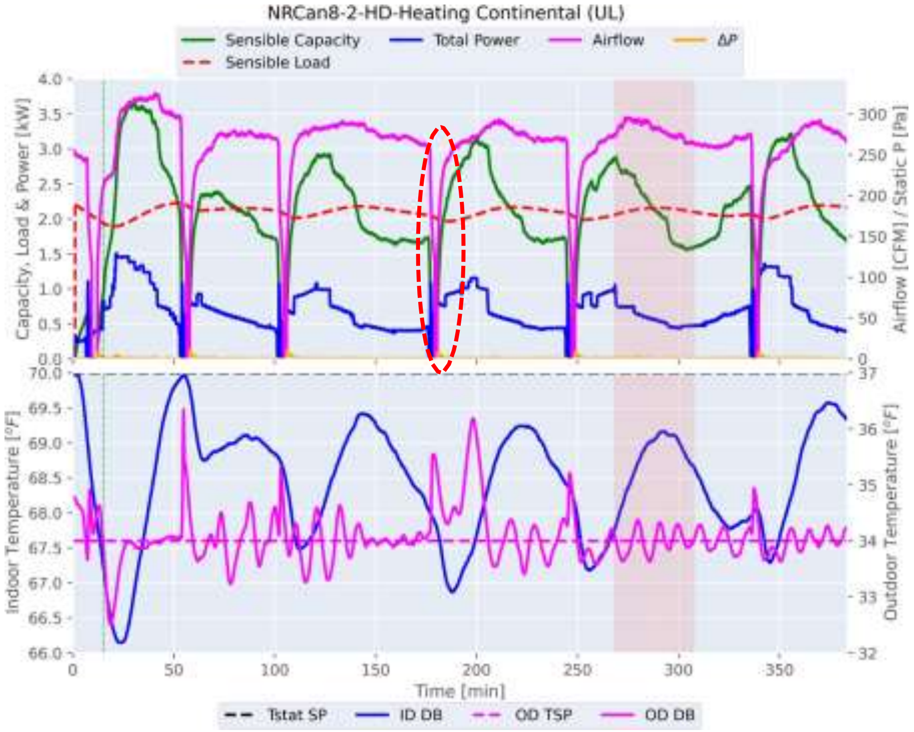


Figure 112. NRCan8-2 (UL) performance and temperature variations for heating continental test interval HD (34°F)

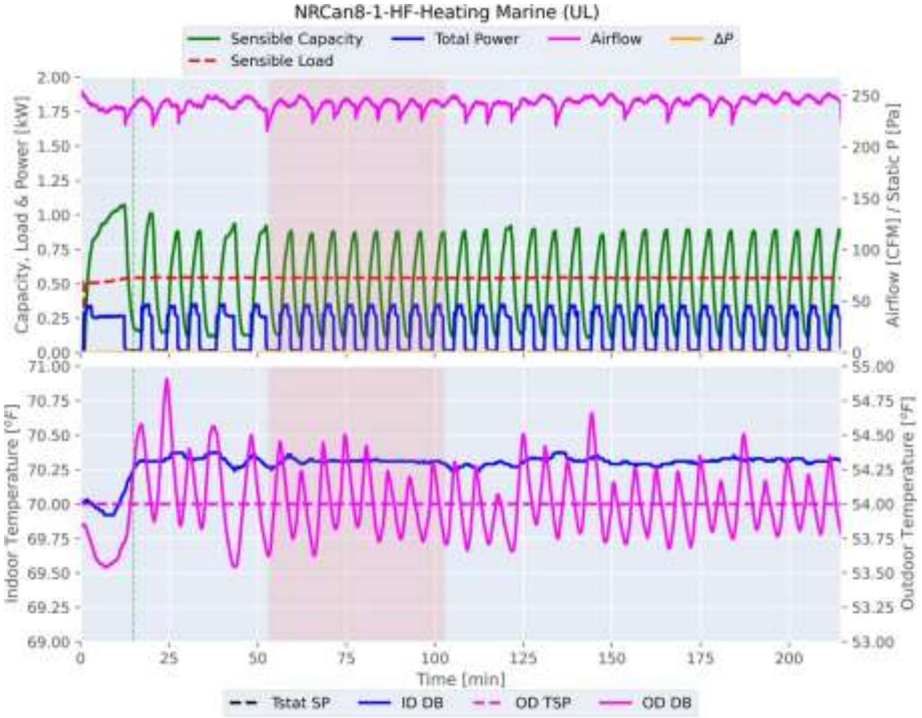


Figure 113. NRCan8-1 (UL) performance and temperature variations for heating marine test interval HF (54℉)

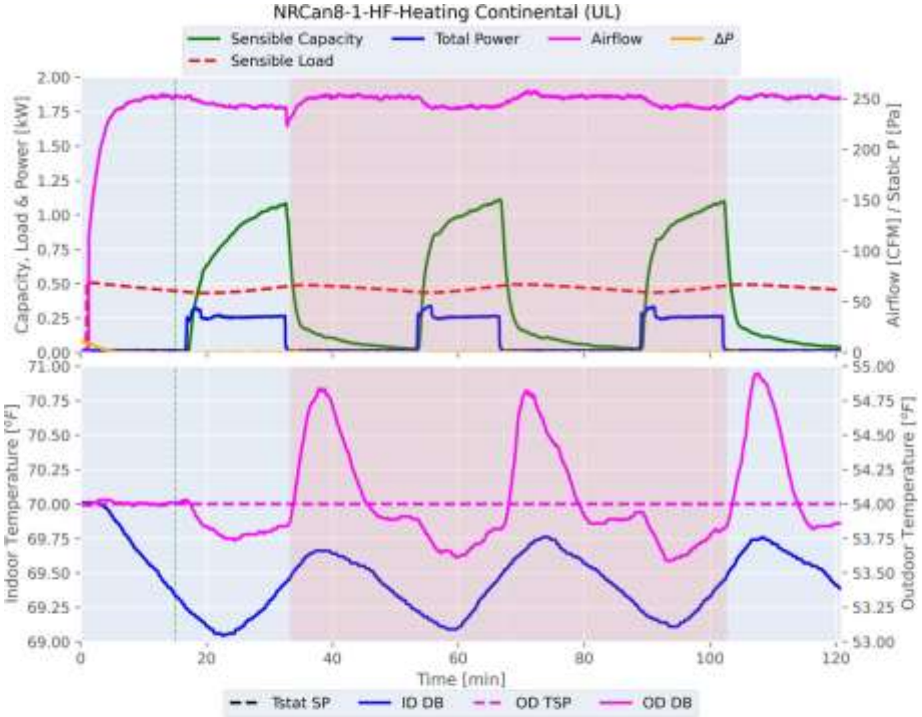


Figure 114. NRCan8-1 (UL) performance and temperature variations for heating continental test interval HF (54℉)

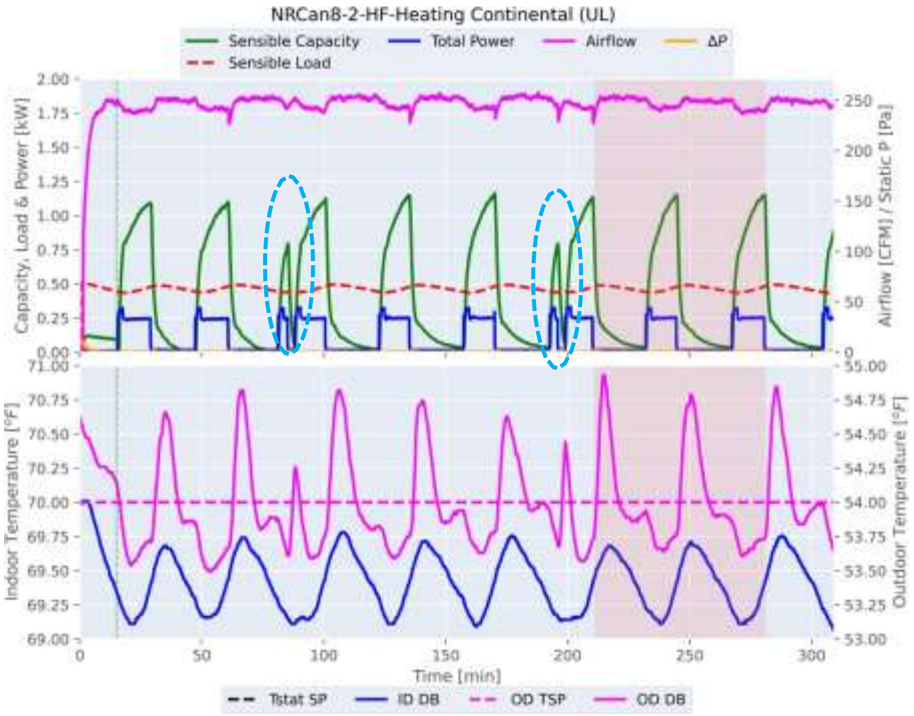


Figure 115. NRCan8-2 (UL) performance and temperature variations for heating continental test interval HF (54℉)

Figure 116 and Figure 117 show NRCan8 estimated heating SCOP comparison for different climate zones along with the maximum differences and average SCOP with a 95% CI and STD based on two sets of tests. Overall test unit showed good repeatability in heating tests with differences in SCOP varying from 0% to 0.8% and STD from ±0% to ±0.4%. Table 19 shows the overall SCOP repeatability results summary with the range and mean of statistical parameters. Overall test unit showed good repeatability with comparatively better heating mode results compared to the cooling mode.

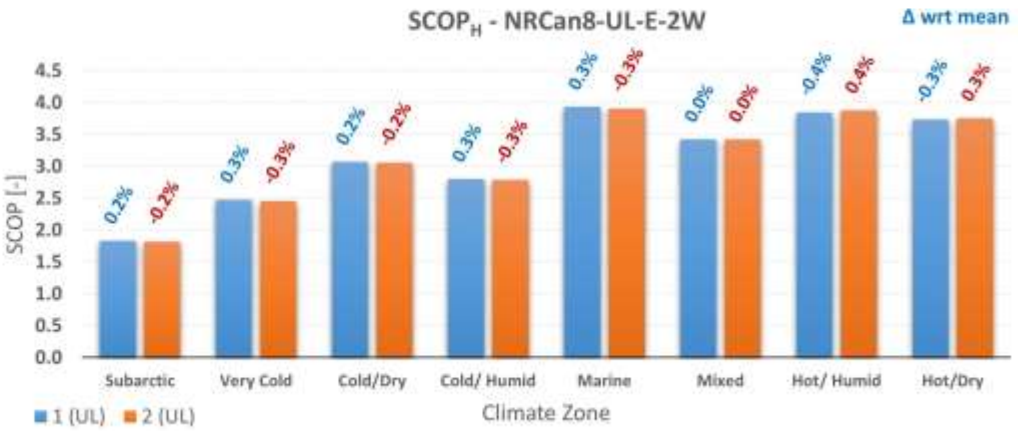


Figure 116. NRCan8 (UL) heating SCOP comparison in different climate zones for different tests

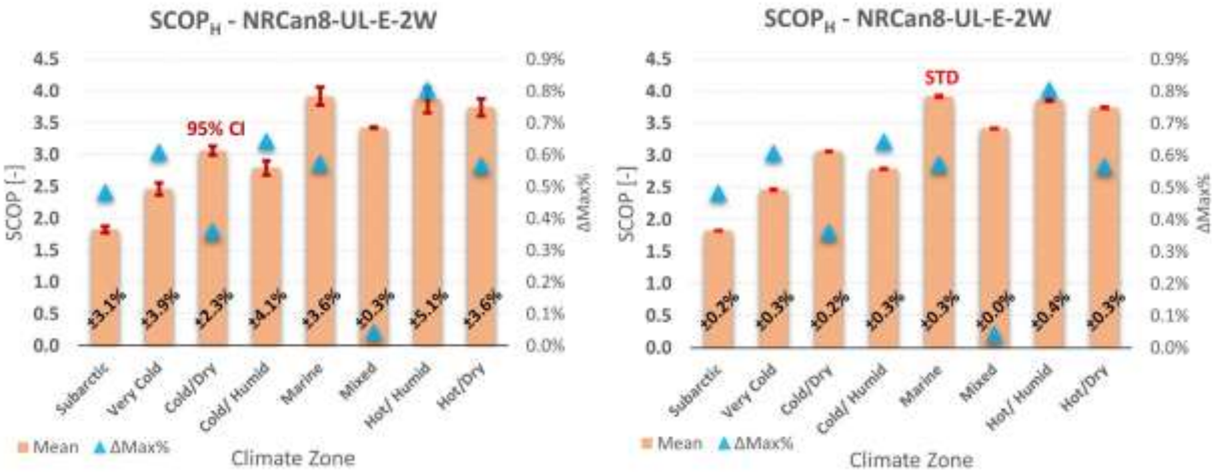


Figure 117. NRCan8 (UL) heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

Table 19. NRCan8-UL-E-2W SCOP repeatability summary results

Metric	No. of Tests	∆SCOP (Max-Min) [%]			SCOP 95% CI [%]			SCOP STD [%]		
		Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
SCOP$_C$	2	0.6%	1.7%	1.1%	3.7%	10.7%	7.0%	0.3%	0.8%	0.6%
SCOP$_H$	2	0.0%	0.8%	0.5%	0.3%	5.1%	3.2%	0.0%	0.4%	0.3%

4.7 NRCan10 (UL)

In this section, NRCan10 unit EXP07 repeatability assessment analysis results are presented based on two sets of test results at the UL lab in dataset NRCan10-UL-E-2W. It is a 2-ton split-type ducted variable-speed heat pump.

4.7.1 Cooling Mode

Figure 118 and Figure 119 show the comparison of test unit performance and operation mode during convergence for cooling dry coil and humid coil test intervals, respectively, with possible issues related to convergence criteria and inconsistent unit dynamic response. At the same test conditions, similar unit operation modes were observed during the convergence period except in the 95°F ambient temperature dry coil test interval, where the unit cycled on/off in test #3 and operated in variable speed hybrid mode in test #4. Figure 120 shows the mean COP of repeated tests for dry and humid coil test intervals with STD and the absolute COP percentage difference. Overall reasonable repeatability in test results was observed. For dry coil test intervals, the absolute difference in COP between the two tests varies from 1.2% to 5.2% and STD from ±0.6% to ±2.6%. Similar repeatability was observed in humid coil test intervals, with COP difference varying from 0 % to 4.6% and STD from ±0% to ±2.3%.

For dry coil test conditions, the largest differences in COP of 5.2% and 4.7% were observed in 77°F and 95°F ambient temperatures intervals, respectively, where better performance was measured in test #3 compared to #4. At 77°F outdoor temperature, the test unit cycled on/off with similar dynamic behavior in both tests and test unit performance did not converge. The difference could be mainly due to some variation in unit dynamic response with its embedded controls and thermostat sensing dynamics in load-based tests. However, at 95°F ambient temperature, the test unit cycled on/off in test #3 and operated in variable speed mode with compressor speed modulation in test #4 as shown in Figure 121. This resulted in different performance, but interestingly better in test #3 where the unit was cycling on/off. One more thing to notice from Figure 121 is that the convergence period did not capture complete compressor modulation cycles, thus not capturing test unit representative performance for this interval. This should be addressed by updating the convergence criteria for cases similar to this.

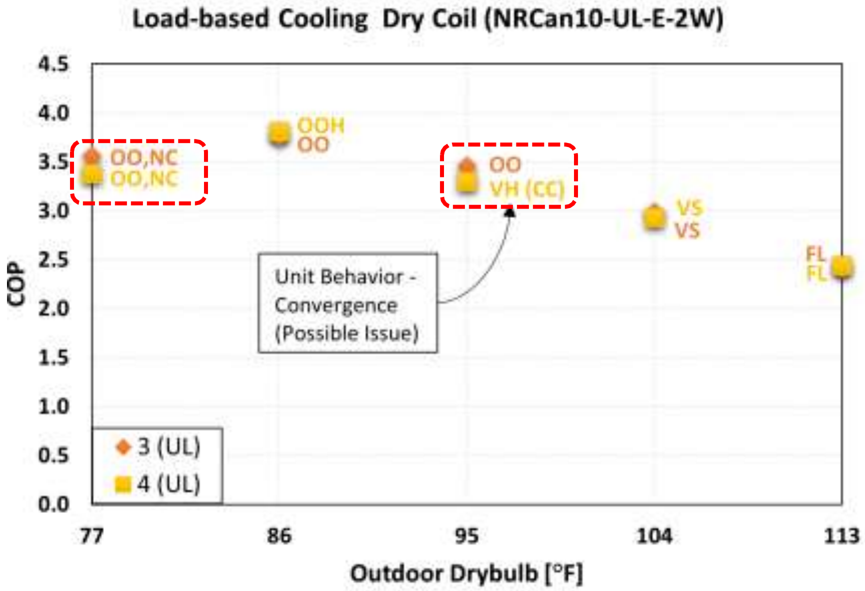


Figure 118. NRCan10 (UL) cooling dry coil test intervals COP for different tests with unit behavior during convergence

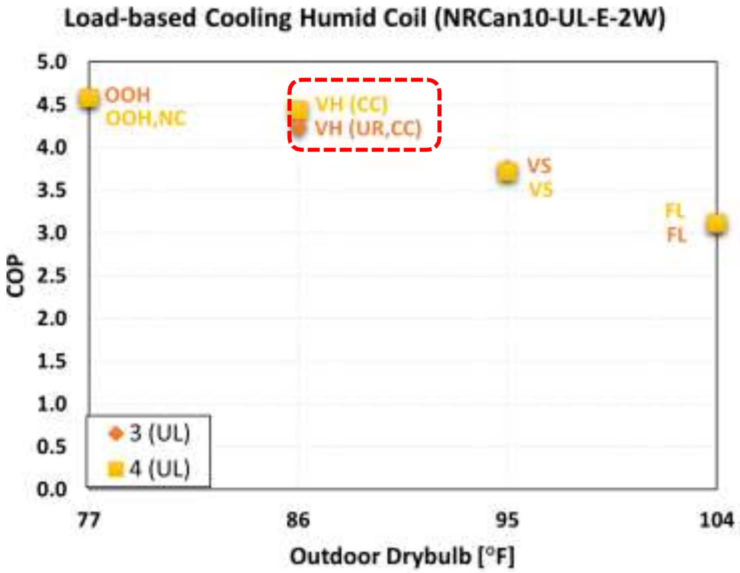


Figure 119. NRCan10 (UL) cooling humid coil test intervals COP for different tests with unit behavior during convergence

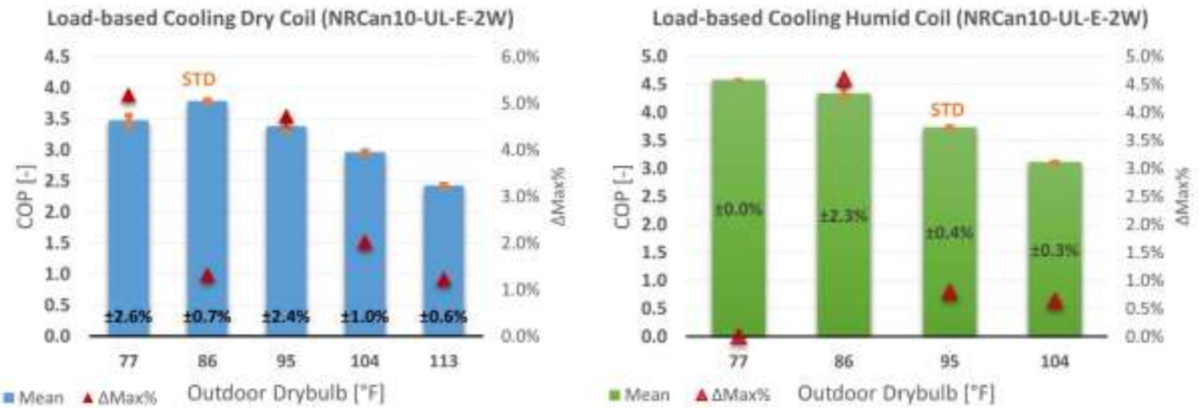


Figure 120. NRCan10 (UL) cooling dry coil and humid coil test intervals mean COP with STD and maximum difference among repeated tests

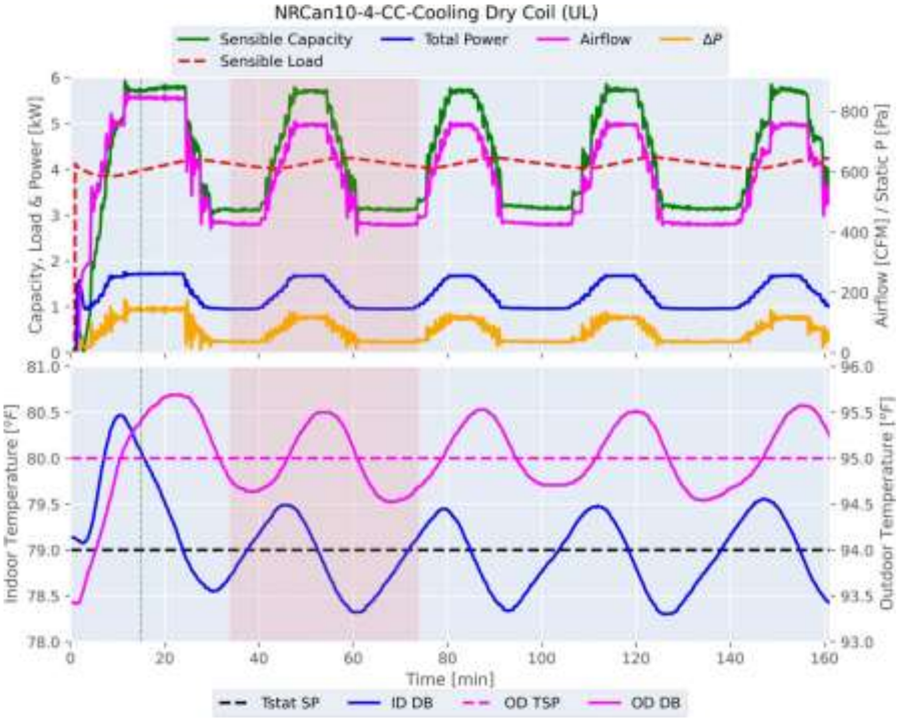


Figure 121. NRCan10-4 (UL) performance and temperature variations for cooling dry coil test interval CC

For humid coil test intervals, the largest difference in COP of 4.6% was observed at 86°F outdoor temperature test conditions, where the unit was operating in variable speed hybrid mode (i.e. compressor speed modulation) during the convergence period as shown in Figure 122 and Figure 123 for test #3 and #4, respectively. The variation in performance is mainly due to the difference in the part of the dynamic response captured during the convergence period between the two tests. Also, both plots show the possible issue with convergence criteria for the case when the compressor is modulating without cycling off, as discussed above for the dry coil interval. Figure 122 also shows that even on running the test for around 6 hours, a steady periodic response was not reached which could be due to the test unit's inconsistent dynamic response with its integrated controller and thermostat. This could also contribute to possible repeatability and reproducibility issues in load-based tests and might not be addressed with updating test methodology as it is more of a test unit dynamic response issue.

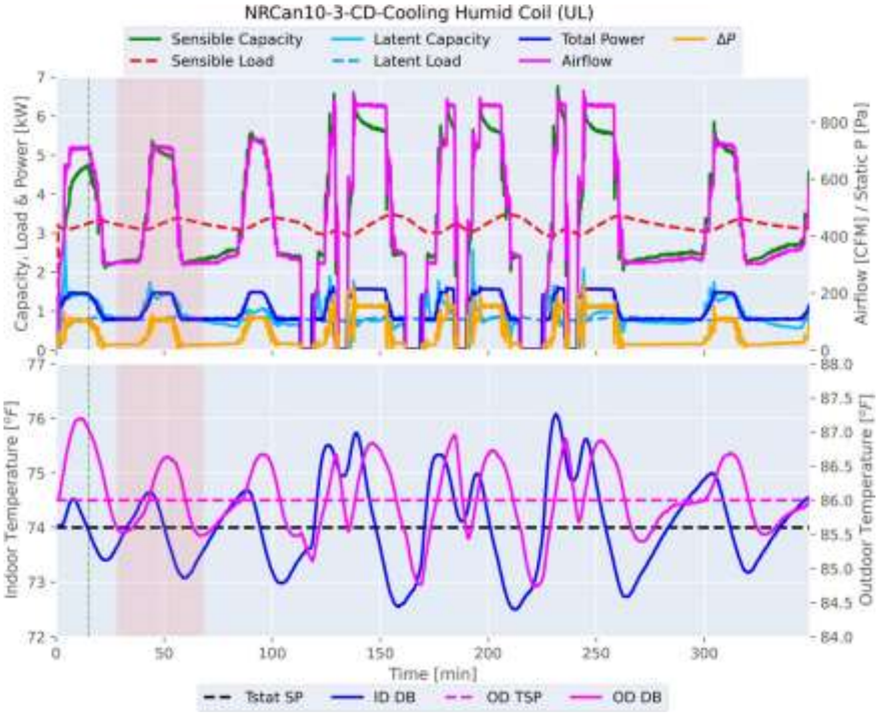


Figure 122. NRCan10-3 (UL) performance and temperature variations for cooling humid coil test interval CD (86°F)

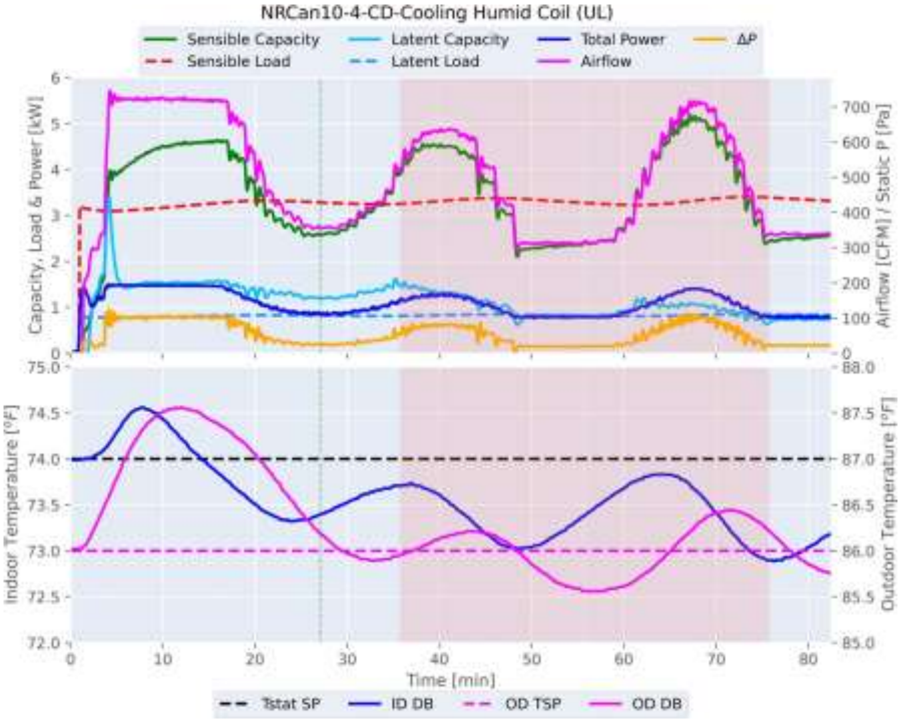


Figure 123. NRCan10-4 (UL) performance and temperature variations for cooling humid coil test interval CD (86°F)

Figure 124 and Figure 125 show the estimated cooling SCOP comparison for NRCan10 across different climate zones between two tests. Overall test unit showed good repeatability with maximum differences in SCOP between two tests varying from 1.7% to 2.7% and STD from ±0.9% to ±1.3% across different climate zones. Slightly better repeatability was observed in climate zones that utilize humid coil test interval results in SCOP estimation.

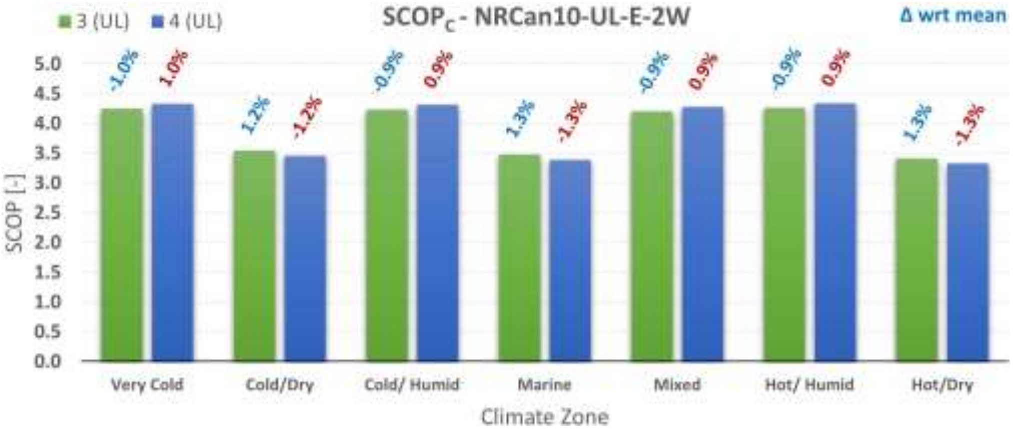


Figure 124. NRCan10 (UL) cooling SCOP comparison in different climate zones for different tests

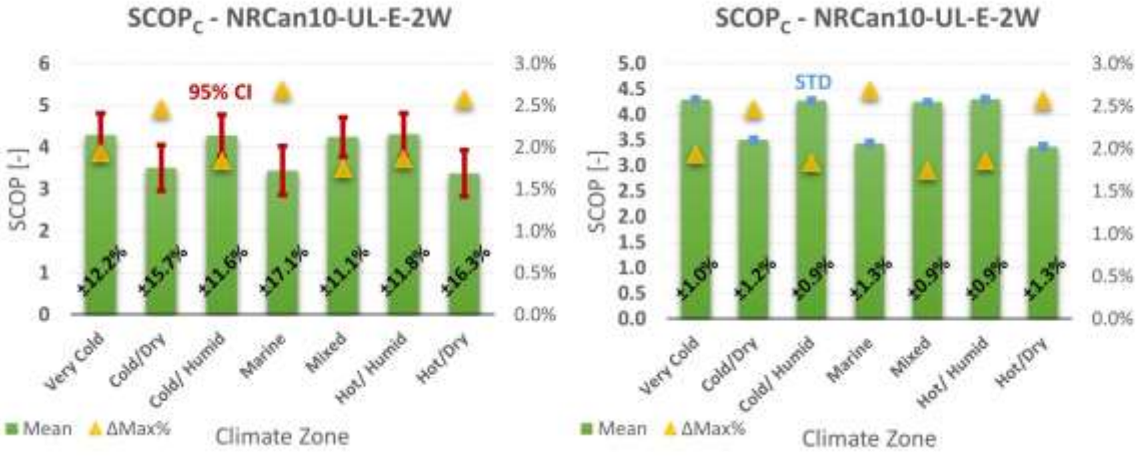


Figure 125. NRCan10 (UL) cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

4.7.2 Heating Mode

Figure 126 and Figure 127 show test unit performance comparisons between two repeated tests in heating continental and marine test intervals, respectively. In repeated tests across different test intervals similar unit operation modes were captured during the convergence period except for continental test interval HB at 5°F outdoor temperature. Figure 128 shows the mean COP of repeated tests for continental and marine test intervals with STD and the absolute COP percentage difference. Overall, the test unit showed good repeatability in heating tests except for HB continental test interval. For marine test intervals, the absolute difference in COP varies from 0.3% to 1.9% and STD from ±0.1% to ±1%. Further, for continental test intervals excluding HB interval, the absolute difference in COP varies from 0% to 2.3% and STD from ±0% to ±1.1%, and in HB interval a larger difference of 10.1% in COP was observed. This large difference is mainly due to the different dynamic responses of UUT in two tests for this full-load test interval at 5°F outdoor temperature as shown in Figure 129 and Figure 130 for tests #4 and #3, respectively. In test #4, the unit operated full-out without going into defrost; whereas, in test #3 unit went into defrost at regular time intervals. This resulted in better performance in test #4 compared to test #3. This difference in defrost behavior could be due to a difference in outdoor humidity, but during the convergence period shown in both tests, the average outdoor wet-bulb temperature was similar. So, this difference could be due to inconsistent unit dynamic response based on its control response.

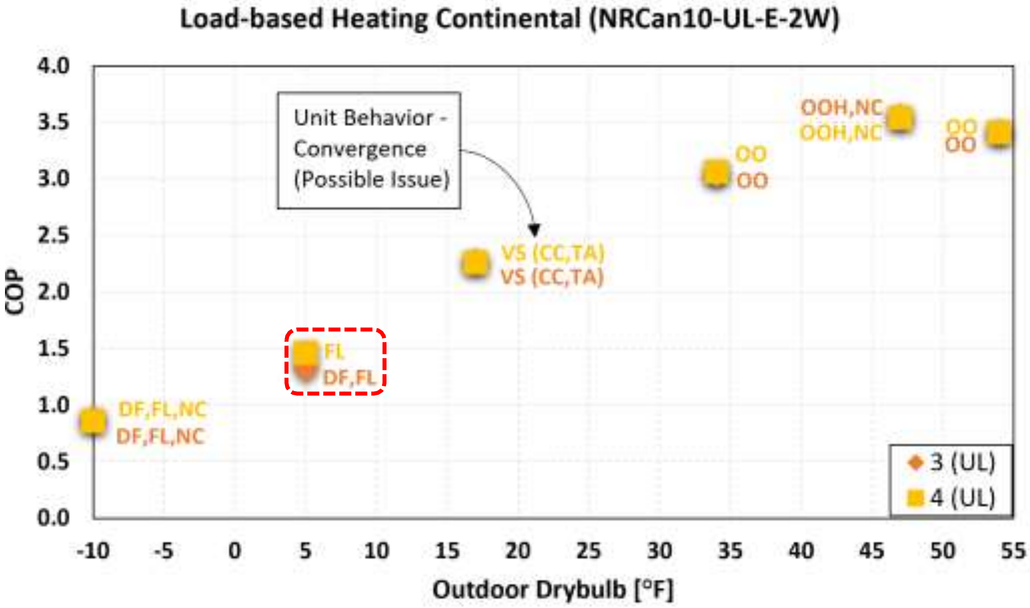


Figure 126. NRCan10 (UL) heating continental test intervals COP for different tests with unit behavior during convergence

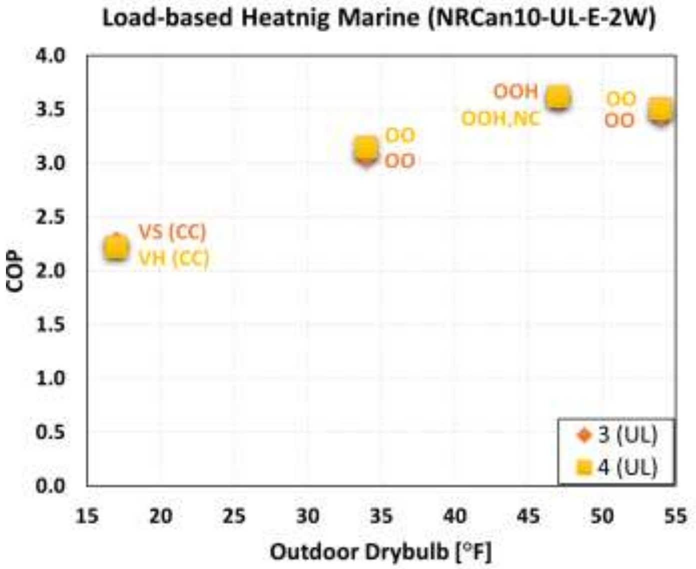


Figure 127. NRCan10 (UL) heating marine test intervals COP for different tests with unit behavior during convergence

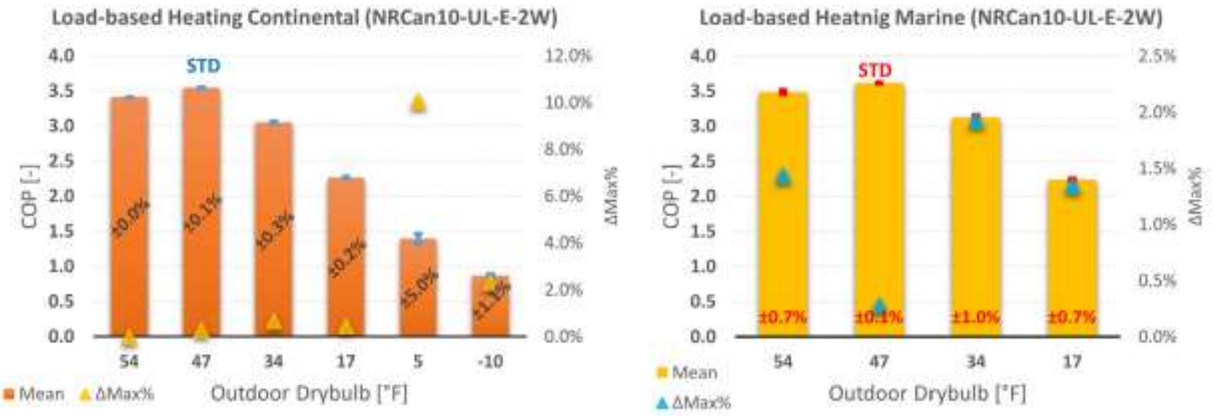


Figure 128. NRCan10 (UL) heating continental and marine test intervals mean COP with STD and maximum difference among repeated tests

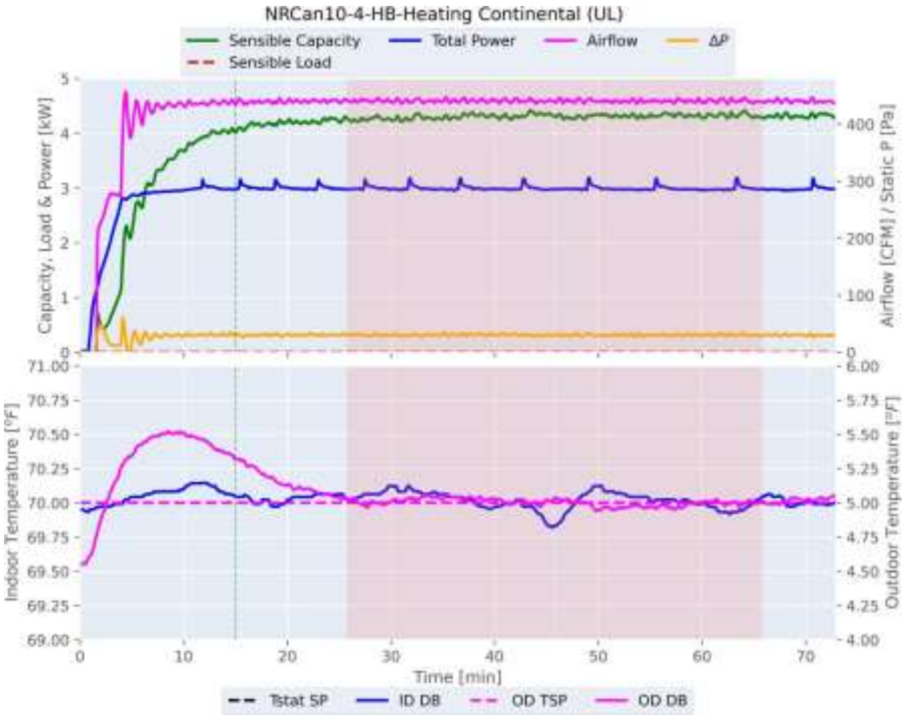


Figure 129. NRCan10-4 (UL) performance and temperature variations for heating continental test interval HB (5°F)

Moreover, some possible issues related to convergence criteria and testing approach in test intervals where units utilize defrost were observed. For example, in the 17°F ambient temperature marine test interval for test #4, as shown in Figure 131, the test unit was going into defrost, however, the convergence was found between two defrosts in a 40 min window and did not account for defrost, thus not capturing representative performance for this test interval. A similar issue was observed in continental test intervals at 34°F outdoor temperature. The testing approach issue here means that test was not run longer to capture multiple complete defrost cycles.

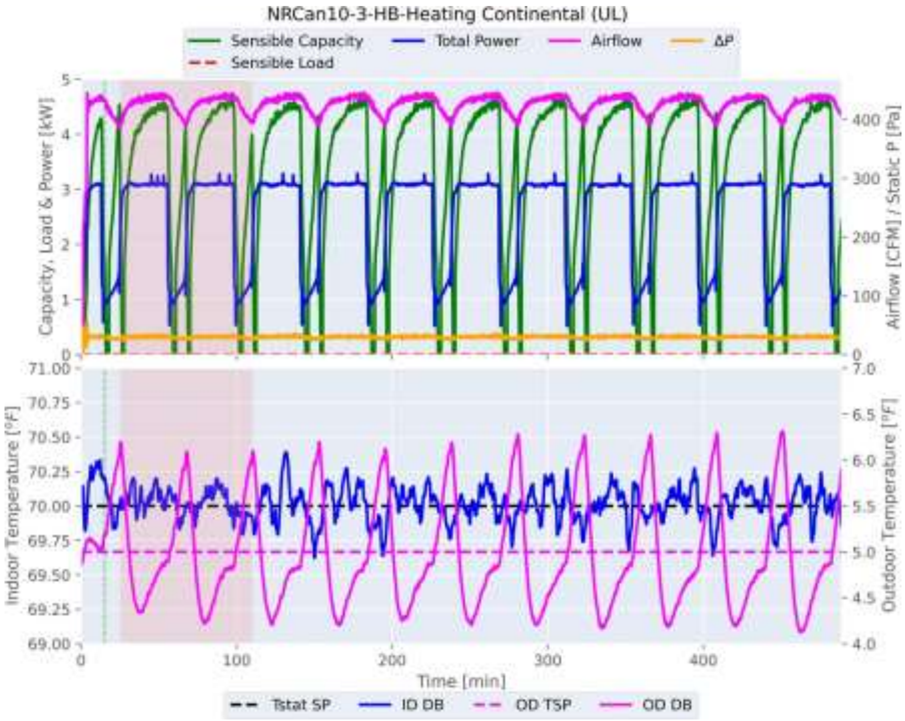


Figure 130. NRCan10-3 (UL) performance and temperature variations for heating continental test interval HB (5℉)

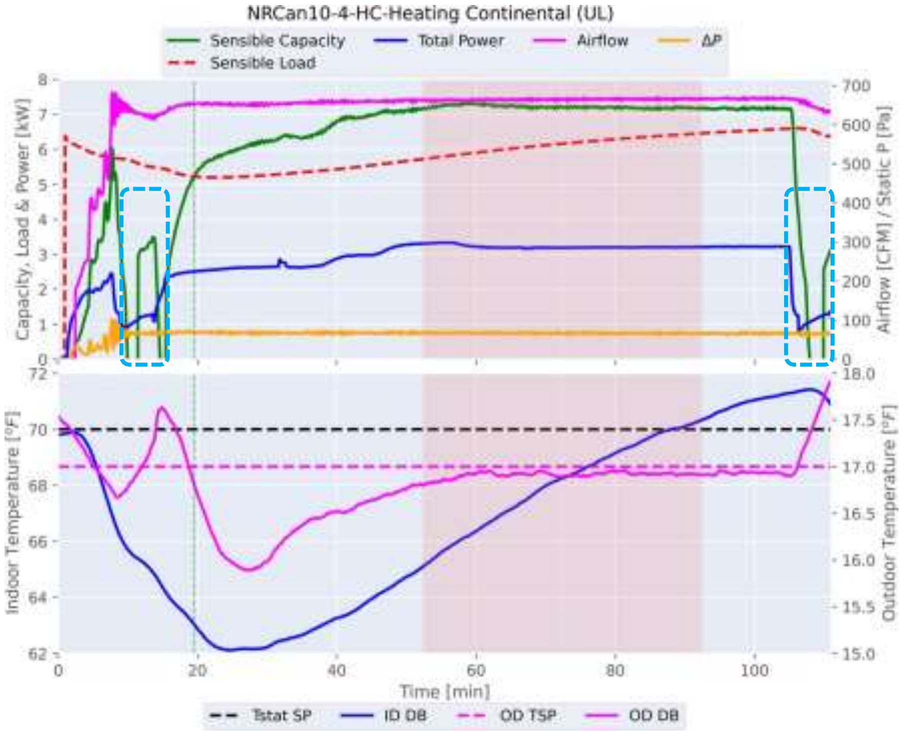


Figure 131. NRCan10-4 (UL) performance and temperature variations for heating continental test interval HC (17℉)

Further, a possible issue with test facility control related to external static pressure control was also observed in some heating test intervals as shown as an example in Figure 132 during heating marine test interval at 34°F outdoor temperature. In the 2nd and half of the 3rd cycle, large fluctuations in airflow and external static pressure (ΔPP) can be seen which possibly seems related to the code-tester booster fan control. Here a similar behavior was also observed in test #4 under the same test conditions, thus not affecting repeatability significantly. However, this could affect load-based test results' reproducibility between different test facilities. It is recommended to further look into this and update the testing approach, if necessary, maybe defining tolerance on external static pressure controllability and acceptable bounds for fluctuations.

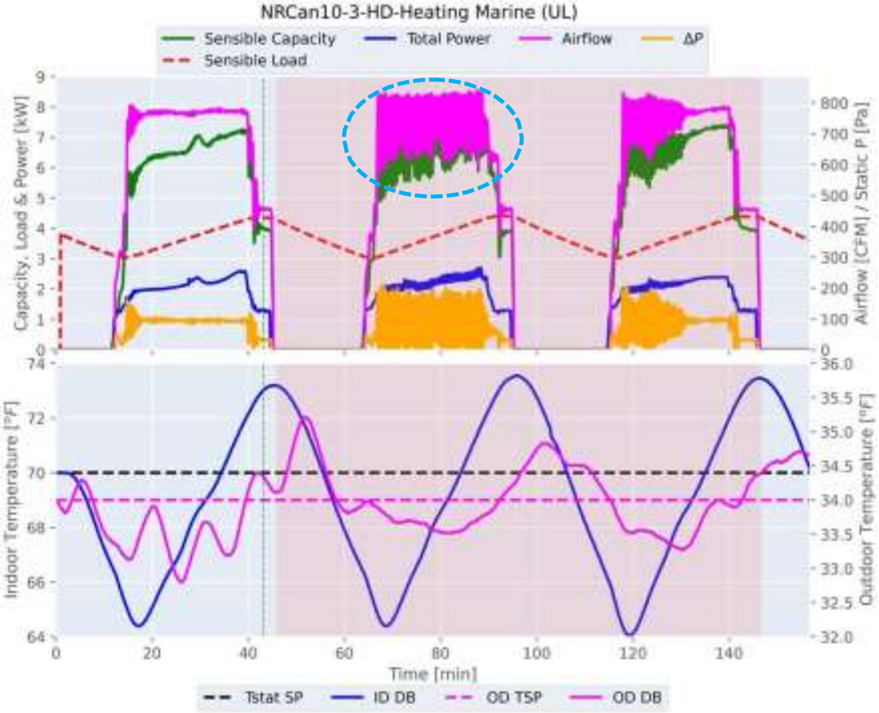


Figure 132. NRCan10-3 (UL) performance and temperature variations for heating marine test interval HD (34°F)

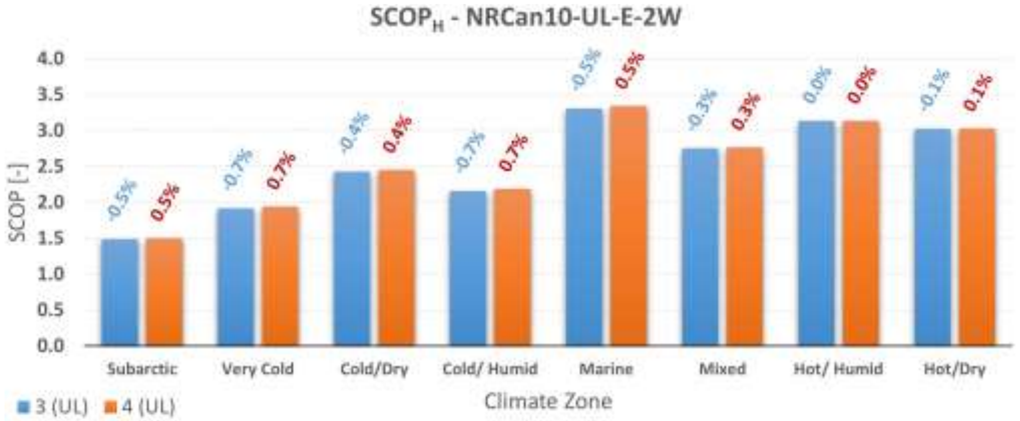


Figure 133. NRCan10 (UL) heating SCOP comparison in different climate zones for different tests

Figure 133 and Figure 134 show a comparison of estimated heating SCOP based on repeated tests in dataset NRCan10-UL-E-2W for different climate zones along with the maximum differences and average SCOP with a 95% CI and STD. Overall NRCan10 unit showed good repeatability in heating with differences in SCOP varying from 0.1% to 1.4% and STD from ±0% to ±0.7%. Table 20 shows the overall SCOP repeatability results summary with the range and mean of statistical parameters for dataset NRCan10-UL-E-2W. Overall test unit showed good repeatability with comparatively better heating mode results compared to cooling mode.

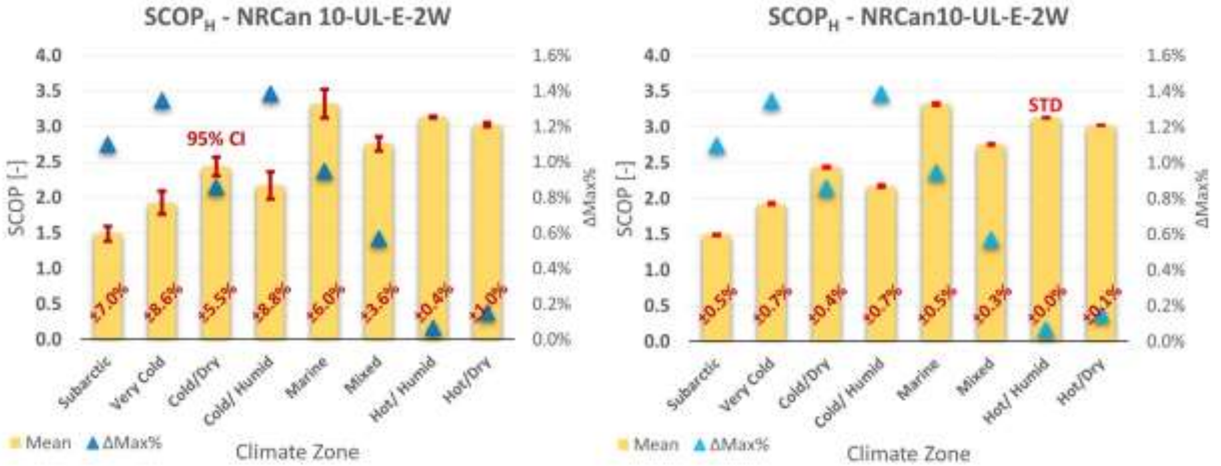


Figure 134. NRCan10 (UL) heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

Table 20. NRCan10-UL-E-2W dataset SCOP repeatability summary results

Metric	**No. of Tests**	**ΔSCOP (Max-Min) [%]**			**SCOP 95% CI [%]**			**SCOP STD [%]**		
		Min	***Max***	***Mean***	***Min***	***Max***	***Mean***	***Min***	***Max***	***Mean***
$SCOP_C$	2	1.7%	2.7%	2.2%	11.1%	17.1%	13.7%	0.9%	1.3%	1.1%
$SCOP_H$	2	0.1%	1.4%	0.8%	0.4%	8.8%	5.1%	0.0%	0.7%	0.4%

5 EXP07 Reproducibility Assessment

In this section, CSA EXP07:19 reproducibility assessment is presented based on the test results of two different heat pump models, NEEA4 and NEEA7(B) (NEEA7& NEEA7B) tested at UL and PG&E labs based on two different datasets defined in Table 3. In the subsections below cooling and heating analysis results for each test unit are presented. First, each test interval COPs are compared along with the observed test unit behavior during convergence and any possible issue flags as per Table 10 and Table 11 for a quantitative and qualitative comparison. In addition, some key test intervals plots showing the test unit performance with time are presented to understand the root cause of observed differences among different tests in the same or different labs. Finally, the estimated seasonal coefficient of performance (SCOP) based on test unit measured performance in different tests are compared to understand the overall variability in seasonal performance metric to assess EXP07 reproducibility.

5.1 NEEA7(B)

In this section, EXP07 reproducibility assessment is presented for NEEA7(B), 3-ton split-type ducted variable-speed heat pump, based on five sets of test results from UL and PG&E lab in dataset NEEA7(B)-UL&PGE-E-5R. Three tests (P2, P3, and #11) were performed at UL, and two (#2 and #4) were performed at PG&E. First, test unit NEEA7 performance was measured in two back-to-back tests P2 and P3 at UL around October 2019. Then unit was shipped to PG&E for further testing; however, some issues occurred with the unit while setting it up at PG&E. Then a different unit but the same model (NEEA7B) was used for performing two back-to-back tests, #2 and #4, at PG&E from around April - July 2021. Then NEEA7B was shipped to UL and final test #11 was performed in the August – September 2021 timeframe. Below, cooling mode results are presented first followed by heating mode reproducibility assessment results for NEEA7(B).

5.1.1 Cooling Mode

Figure 135 shows the test unit measured COP at different cooling dry coil intervals based on EXP07 for all 5 different tests in dataset NEEA7(B)-UL&PGE-E-5R, along with the unit operation mode (Table 10) during the convergence period and possible issues (Table 11) for each test interval. The test unit cycled on/off at low ambient temperatures with low building load, operated in variable speed mode at intermediate outdoor temperatures, and failed to meet the building load at higher outdoor temperatures, where a full-load test was performed. Similar unit operation modes were observed during convergence in different tests at similar test conditions except at 95°F outdoor temperature test intervals. In this test interval, the test unit operated in variable speed steady or hybrid mode at UL, whereas, cycled on/off at PG&E to meet the virtual building load. A possible issue with convergence criteria noted at 95°F and 104°F ambient temperature interval in test #2 is related to the convergence period not capturing the representative performance in the case of test unit operating in variable speed hybrid mode where compressor modulates up and down in a regular pattern without cycling off, as discussed above in repeatability results section 4.5. However, it should be noted that these possible issues do not play a significant role in observed performance differences in this case.

To understand the quantitative variation in unit performance for different tests, Figure 136 shows the comparison of measured COP at different ambient conditions for different cooling dry coil tests along with the percentage difference of each test interval COP with the mean value of 5 tests at the same test conditions. In general, higher COP was measured at UL compared to PG&E lab across cooling dry coil test intervals.

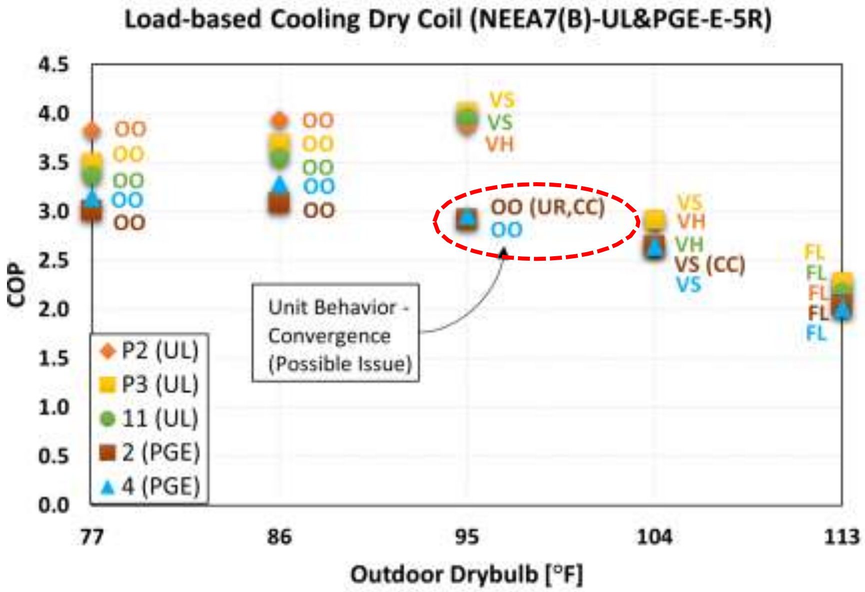


Figure 135. NEEA7(B) cooling dry coil test intervals COP with unit behavior during convergence

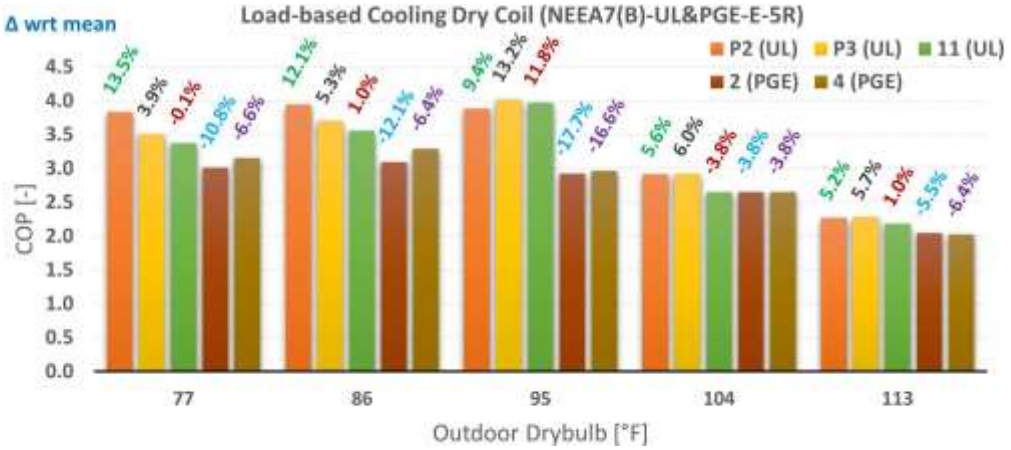


Figure 136. NEEA7(B) cooling dry coil test intervals COP comparison

For 77°F and 86°F ambient temperature intervals, where the unit was cycling on/off in both labs, a large variation in COPs was observed with maximum COP measured in test P2 at UL and minimum in test #2 at PG&E. In different tests, COP from the mean of 5 tests varies from -10.8% to 13.5% and from -12.1% to 12.1% for 77°F and 86°F test intervals, respectively. In both test intervals, the order of tests from maximum to minimum COP is the same which indicates there are possibly additional factors contributing to performance variations other than just random test and measurements uncertainties. Figure 137, Figure 138, and Figure 139 show the test unit performance with time for 77°F dry coil test interval in tests P2, #11, and #4, respectively. In three tests, there was variation in the indoor temperature maintained. During the convergence period, in test P2 (UL) the indoor temperature varied from 76°F to 79.3°F with an average value of 77.6°F, in test #11 (UL) it varied from 77.1°F to 80.7°F with an average of around 79°F, and in test #4 (PG&E) it varied from 75.8°F to 80.5°F with an average of around 78.3°F. This difference in the average indoor temperature maintained as well as the range of indoor temperature variation is possibly due to thermostat sensing dynamics and its offset settings which could be affected by thermostat installation location, test facility air velocity and temperature distribution, and thermostat offset setting approach as per EXP07. It seems like the lower average indoor temperature maintained in test P2 and similarly in test P3 could be due to a difference in thermostat offset settings compared to the other three tests. The indoor temperature controlled based on virtual building response during load-based testing is coupled to the virtual building load. So, variation in indoor temperature affects the virtual building load and vice versa, which affects the overall performance measured during dynamic tests. Also, between the two labs, a difference in test unit rate of change of sensible cooling rate was observed when the compressor turns on and off as can be seen in Figure 138 and Figure 139. At UL measured sensible cooling rate change step response is slower compared to PG&E when the compressor turns on and off, which is possibly due to difference in the thermal mass of measurement system between the unit outlet and supply air measurements as well as due to difference in instrumentation and test setup between two labs and the same lab on unit reinstallation. This dynamic difference affects the virtual building model response, thus indoor temperature variation and overall performance. In the three tests shown below, unit on/off cycling rate varied with cycle time during the convergence period of around 62 min in test P2, 80 min in test #11, and 75 min in test #4. Also, at PG&E, around 150 CFM lower airflow with around 10 Pa higher external static pressure was observed compared to UL lab at the same test conditions. This could be due to a difference in external static pressure control as per the defined virtual duct model in EXP07 based on the design airflow rate and also possibly due to fan airflow settings during the test unit setup. So, at 77°F and 86°F cooling dry coil test intervals, the variability in performance is likely due to difference in test unit dynamic response with its integrated controller, thermostat offset and its sensing dynamics, external static pressure control, test setup, and instrumentation in addition to measurement and dynamic test uncertainties and differences in test facility control to maintain test conditions.

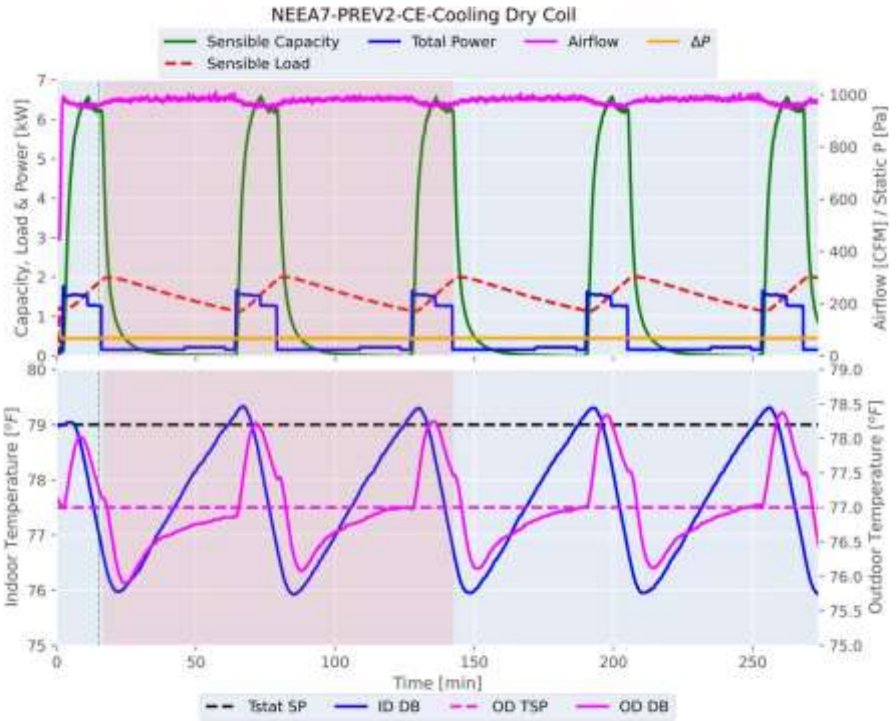


Figure 137. NEEA7-P2 (UL) performance and temperature variations for cooling dry coil test interval CE (77℉)

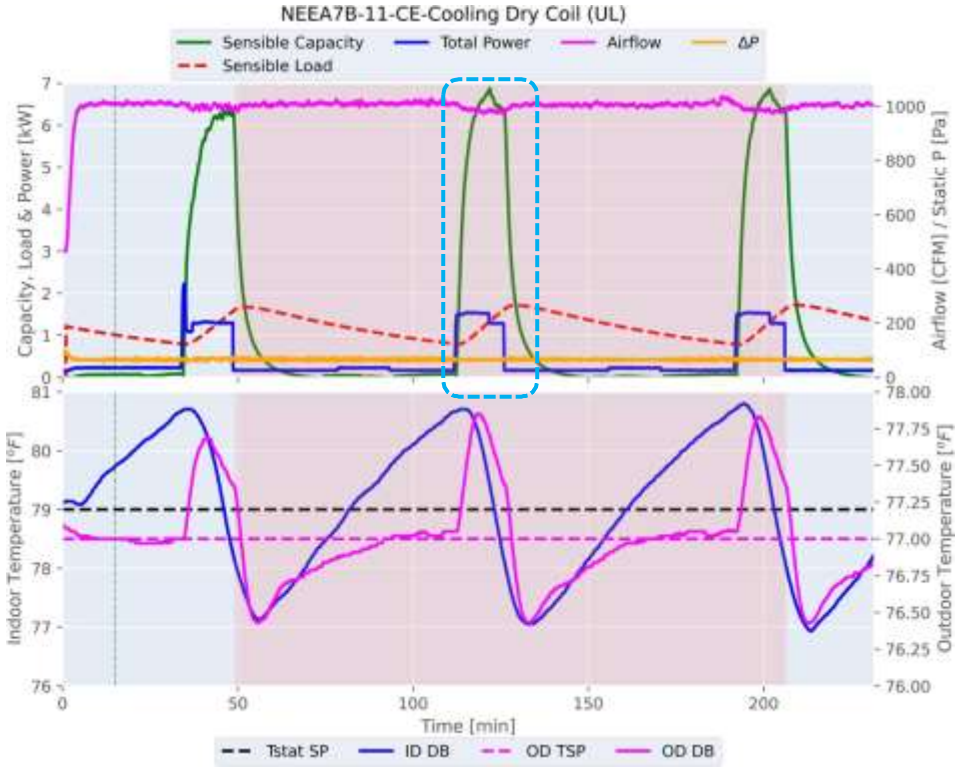


Figure 138. NEEA7B-11 (UL) performance and temperature variations for cooling dry coil test interval CE (77℉)

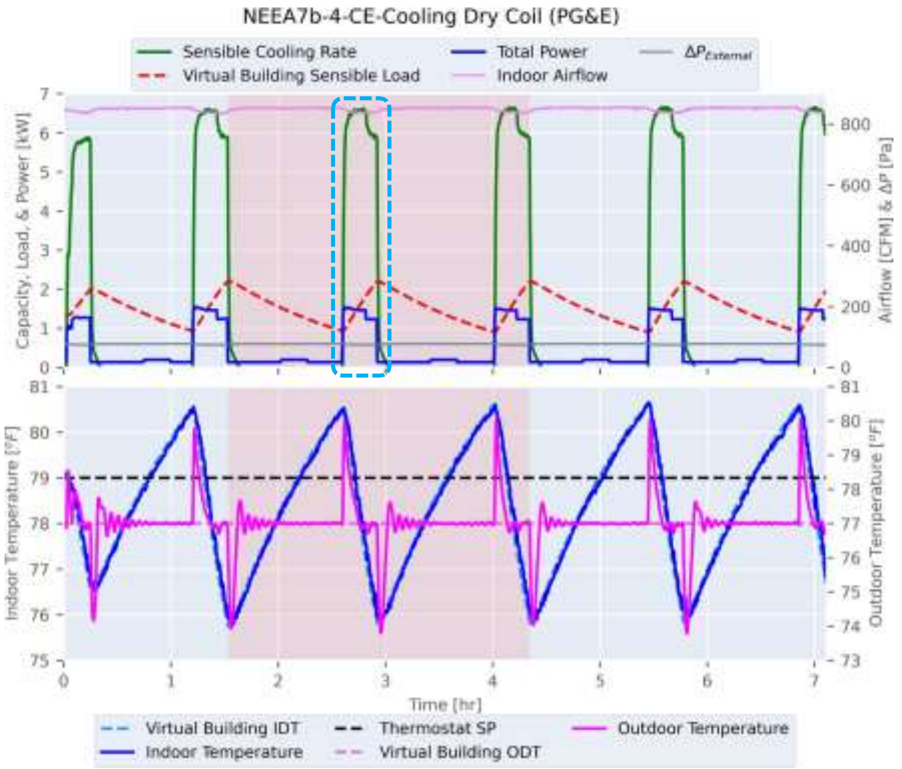


Figure 139. NEEA7B-4 (PG&E) performance and temperature variations for cooling dry coil test interval CE (77°F)

At 95°F outdoor temperature dry coil test interval, the test unit showed good repeatability in measured performance at UL and PG&E; however, quite lower COP was measured at PG&E compared to UL with the difference in COP from mean of 5 tests varying from -17.7% to 13.2%, lowest COP in test #2 at PG&E and highest in test P3 at UL. This large difference in performance at the two labs is mainly due to different test unit dynamic behavior under the same test conditions, where the unit operated in variable speed mode at UL as shown in Figure 140 for test P3 and cycled on/off at PG&E as shown in Figure 141 for test #4, resulting in a lower performance at PG&E compared to UL. This difference in dynamic behavior could be due to variation in thermostat sensing dynamics and test unit controller response accordingly. In test P3 at UL, UUT turns on when the indoor temperature is around 79°F and keeps the compressor speed constant with slightly increasing indoor temperature to a maximum of around 79.75°F and then the compressor step down to a lower speed without cycling off as the indoor temperature reaches around 78.5°F and convergence was found in a 40 min window. However, in test #4 at PG&E, UUT turns on when the indoor temperature reaches around 82°F to a similar compressor speed as UL comparing outdoor power consumption. Then as the indoor temperature is quite higher than the thermostat setpoint, the compressor ramps up resulting in a higher cooling rate compared to the building load and a fast rate of change in indoor temperature. Then as the indoor temperature decreases around 77.6°F, the compressor ramps down to a similar minimum compressor speed as the UL test; however, after some time test unit cycled off rather than operating in variable speed mode as UL, possibly due to faster change in indoor temperature at PG&E. Following that similar on/off cycling pattern continues in the PG&E test and convergence was found in two successive on/off cycles. This difference in test unit dynamic response with its integrated controller and thermostat seems to be the main cause of performance variation between the two labs in addition to other factors related to the test unit setup, external static pressure control, test conditions control, and instrumentation with its associated uncertainties.

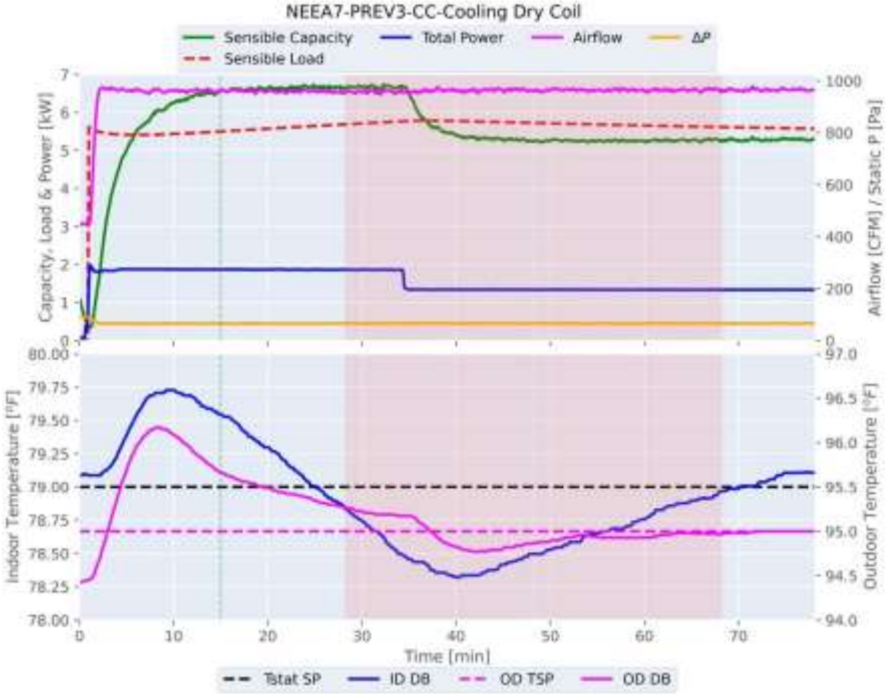


Figure 140. NEEA7-P3 (UL) performance and temperature variations for cooling dry coil test interval CC (95℉)

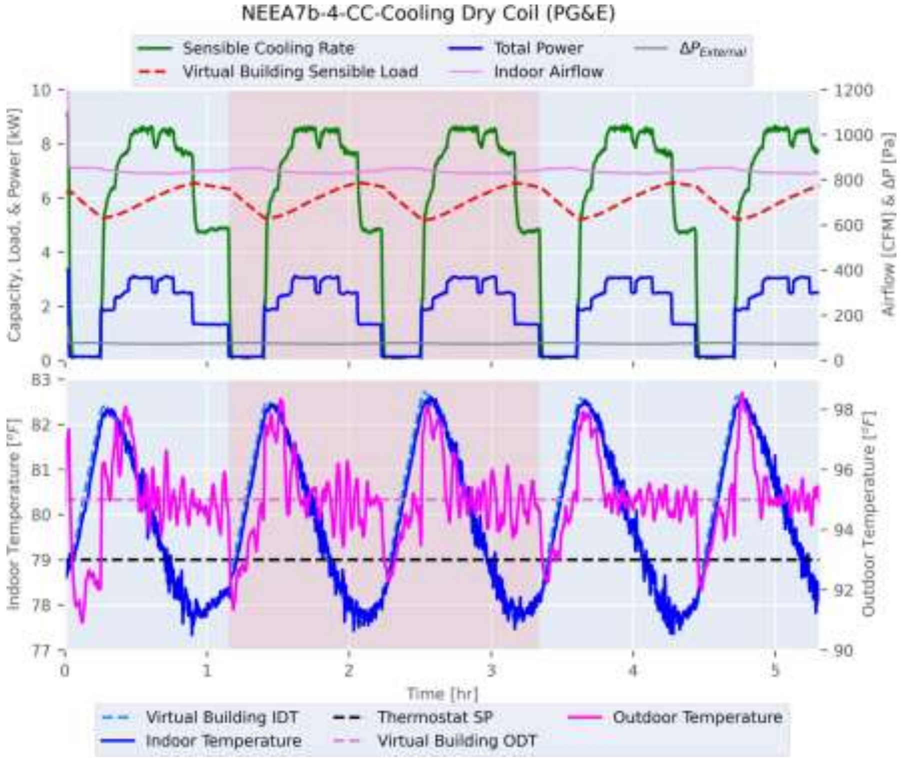


Figure 141. NEEA7B-4 (PG&E) performance and temperature variations for cooling dry coil test interval CC (95℉)

In cooling dry coil test intervals, the lowest differences in performance among 5 tests in two labs were observed for 104°F ambient temperature intervals with differences in COP from the mean of 5 tests varying from -3.8% to 6%. Comparable COPs were observed in tests P2 and P3 at UL with NEEA7 which were higher than the similar COPs observed in other 3 tests, #11 at UL and #2 & #6 at PG&E, with NEEA7B. In this test interval, the test unit operated in variable speed mode to meet the building load. Some of the factors in observed differences between the two labs could be due to variation in test unit dynamic response as well as convergence period at same test conditions. At UL average indoor temperature during convergence was around 79°F; whereas, at PG&E it was higher, 80.4°F in test #2 and 79.7°F in test #4. Also, similar to other intervals, airflow in PG&E tests was lower than observed at UL. The differences in the two test labs without the effect of the test unit's dynamic response with its embedded controller and thermostat could be understood based on the performance differences observed in full-load tests at the 113°F outdoor temperature. In this test interval, relatively better COPs were observed in UL tests compared to PG&E tests with variation in COPs from the mean of 5 tests from -6.4% to 5.7%. These differences are mainly due to variations in test unit setup, instrumentation, and external static pressure control. In UL full-load tests, the airflow was around 1400 CFM with 124 Pa (0.5 wc) external static pressure, whereas, in PG&E tests, airflow was around 1200 CFM with 153 Pa (0.6 wc) external static pressure.

Figure 142 and Figure 143 show the NEEA7(B) performance comparison for cooling humid coil test intervals in 5 sets of tests in two labs along with test unit operation mode during convergence and possible issues for each test interval. At the same test conditions, similar unit operation modes were observed with a convergence criteria issue at the 95°F outdoor temperature test interval. As discussed in section 4.4.1, this convergence criteria issue is related to the convergence period not capturing representative performance when the test unit is operating in variable speed hybrid mode with compressor modulating without cycling off. The largest differences in performance were observed in the 95°F test interval, especially with relatively higher COP in test P2 compared to the other 4 tests with similar COPs. This difference is due to variation in test unit dynamic response as well as the period where convergence was found as shown in Figure 144 for test P2 at UL and in Figure 145 for test #4 at PG&E. In test P2, convergence was found when the unit operating at low compressor speed resulted in around 20% higher COP compared to test #4.

Overall, in cooling humid coil tests, relatively better reproducibility was observed compared to dry coil tests which might seem counterintuitive as additional variability could be expected due to humidity introduction. The better results in humid coil tests could be attributed to additional humidity-related variabilities counteracting some of the factors that cause variations in dry coil test intervals.

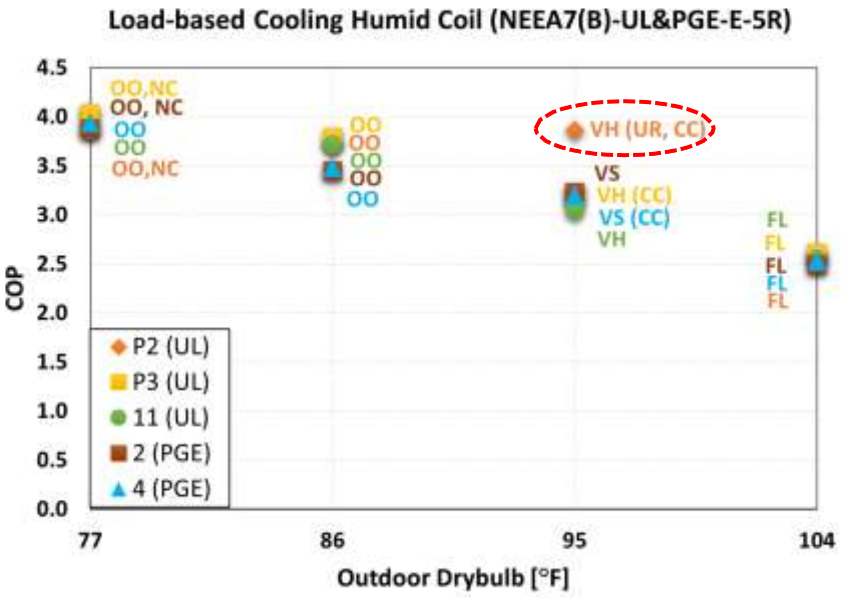


Figure 142. NEEA7(B) cooling humid coil test intervals COP with unit behavior during convergence

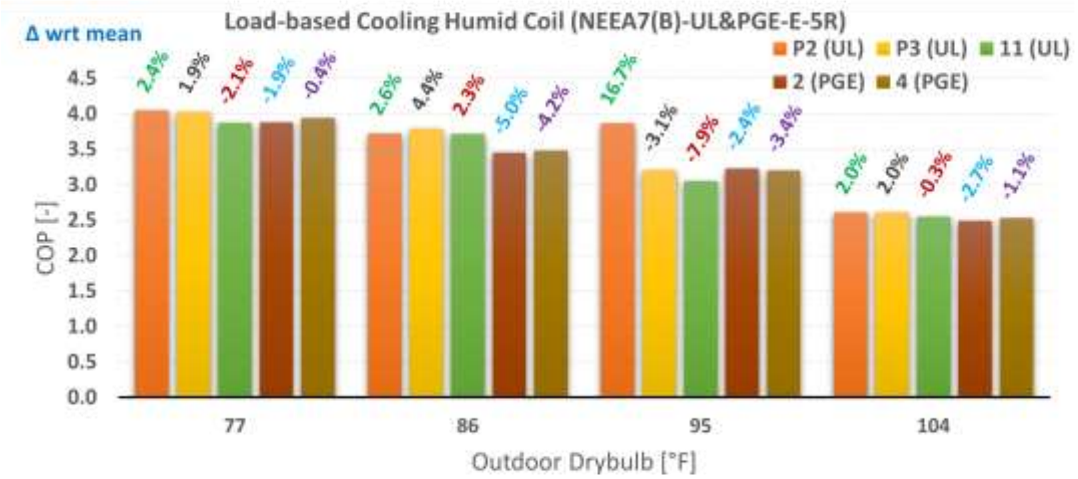


Figure 143. NEEA7(B) cooling humid coil test intervals COP comparison

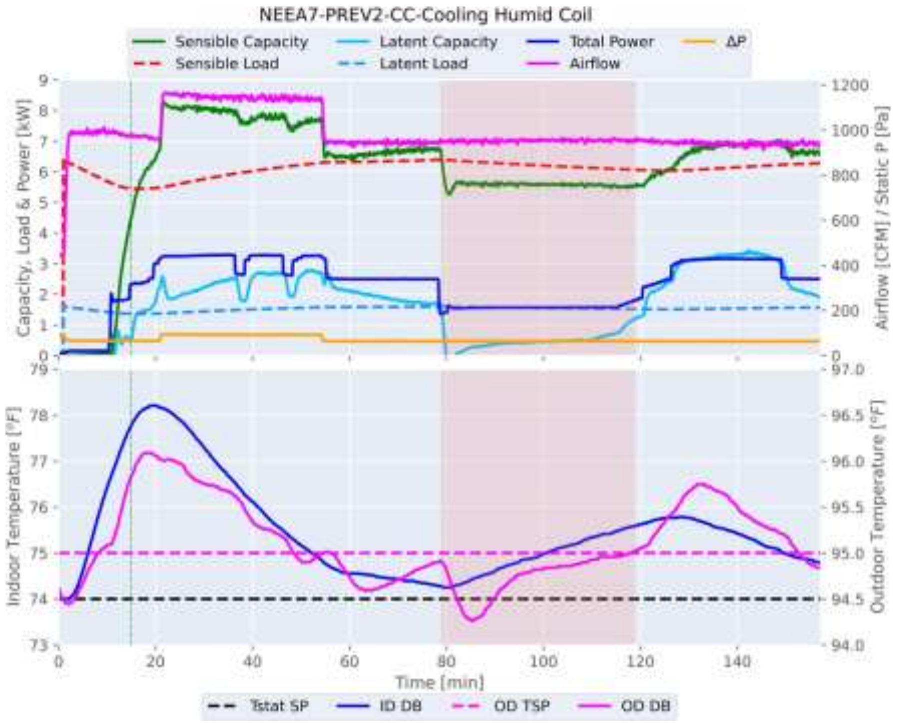


Figure 144. NEEA7-P2 (UL) performance and temperature variations for cooling humid coil test interval CC (95℉)

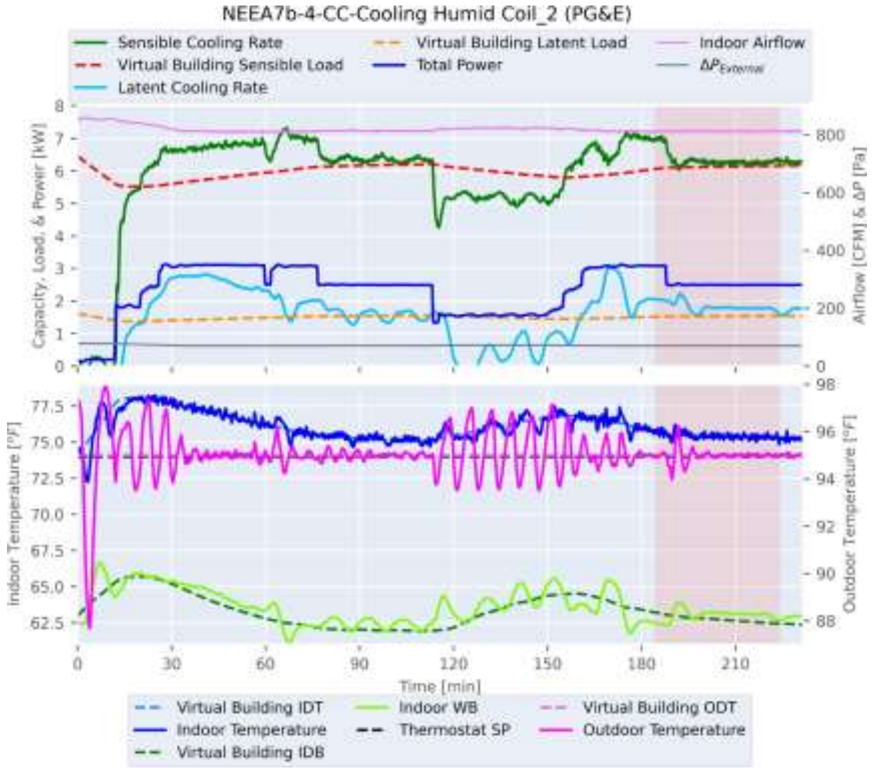


Figure 145. NEEA7B-4 (PG&E) performance and temperature variations for cooling humid coil test interval CC (95℉)

Figure 146 shows with average COP of 5 tests in two labs for cooling dry coil and humid coil test intervals in dataset NEEA7(B)-UL&PGE-E-5R along with STD and the maximum COP difference in percentage on the right vertical axis. Overall poor reproducibility was observed in dry coil test intervals with the maximum difference in COP between different tests varying from 9.8% to 30.9% and STD from ±4.7% to ±14%. Relatively large variations were observed at 77°F, 86°F, and 95°F ambient condition dry coil test intervals. Comparatively better reproducibility was observed in humid coil tests intervals with the maximum difference in COP varying from 4.6% to 24.6% and STD from ±1.8% to ± 8.6%. In general, across all cooling test intervals, relatively better COP was measured in tests at UL compared to PG&E tests at the same test conditions.

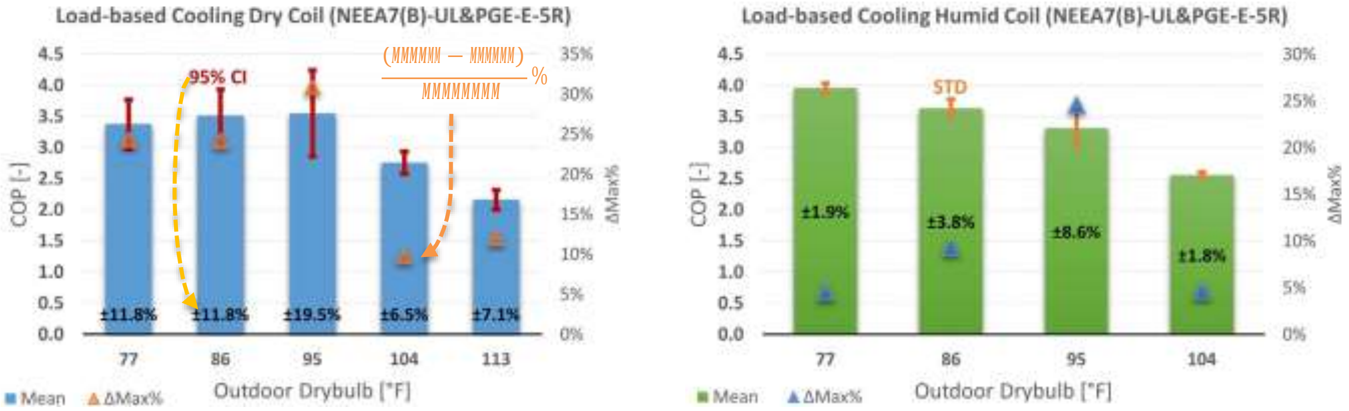


Figure 146. NEEA7(B) cooling dry coil and humid coil test interval COP average and maximum difference for different tests with a standard deviation

Figure 147 and Figure 148 show the comparison of estimated cooling SCOP across different climate zones with maximum percentage differences and average value with a 95% CI and STD for 5 tests in both labs. Corresponding to variations observed in COP across different test intervals, the highest cooling SCOP was observed based on test P2 results at UL and the lowest in test #2 at PG&E. In general, higher SCOPs were estimated based on UL results compared to PG&E test results.

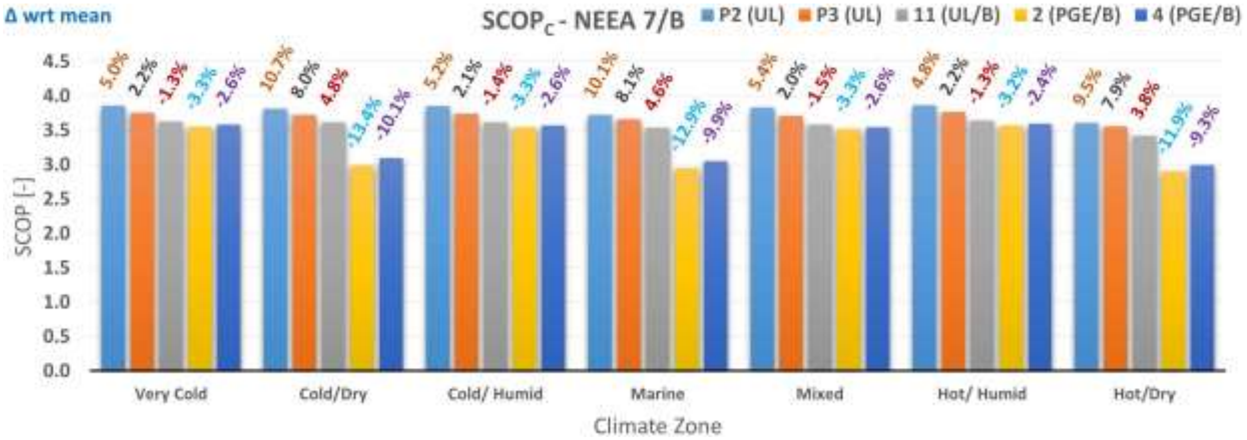


Figure 147. NEEA7(B) cooling SCOP comparison in different climate zones for different tests

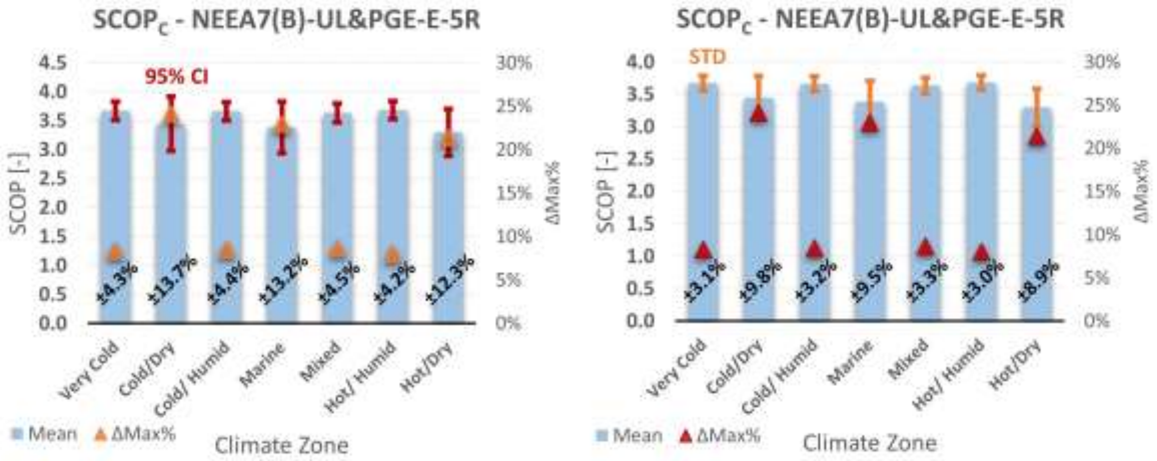


Figure 148. NEEA7(B) cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

For three climate zones utilizing cooling dry coil test results in SCOP estimation, overall poor reproducibility was noted in cooling SCOP with maximum differences varying from 21.4% to 24.1%, 95% CI from ±12.3% to ±13.7%, and STD from ±8.9% to ±9.8%. This is due to a larger variation in cooling dry coil test results among 5 tests as discussed above. On the other hand, for four other climates zones which utilize humid coil test results in SCOP estimation, reasonable reproducibility was observed with maximum differences in SCOP varying from 8% to 8.7%, 95% CI from ±4.2% to ±4.5%, and STD from ±3% to ±3.3%.

5.1.2 Heating Mode

Figure 149 and Figure 150 show NEEA7(B) COP comparison in heating continental test intervals along with test unit operation mode captured during the convergence period and possible issues with test intervals for 5 sets of tests at UL and PG&E in dataset NEEA7(B)-UL&PGE-E-5R. More details on possible issues were discussed above in sections 4.4.2 and 4.5.2.

At higher ambient temperatures of 54°F and 47°F, the test unit cycled on/off in all tests, and in contrast to cooling results, higher COP was measured in tests at PG&E compared to UL test results. In both test intervals, a relatively large variation in measured COP was observed with differences in COP from the mean of 5 tests varying from -7% to 6.4% for 54°F test interval and from -7.3% to 6.2% for 47°F test interval. Figure 151 and Figure 152 show the test unit performance for 54°F test interval (HF) in test P3 at UL and in test #4 at PG&E respectively. The first thing to notice is that similar to cooling results, a lower airflow rate was observed at PG&E compared to UL, however, a slightly higher heating rate was measured in the PG&E test during the cycle *on* period with a similar power consumption rate. Some variation in the heating rate dynamic response was also observed between the two labs when the compressor turns on and off. Further, in test P3 at UL, the average indoor temperature during the convergence period was around 70.5°F with fluctuation of around 5.5°F between the minimum and maximum value. Whereas in test #4 at PG&E, average indoor temperature was maintained at around 71.5°F with fluctuations of around 4.5°F. This difference in indoor temperature response between two labs at the same test conditions is possibly due to variation in test unit dynamic response and thermostat sensing dynamics along with its offset. This variation in test unit dynamic performance with its thermostat and embedded controller resulted in different on/off cycling periods with around 60 min in test P3 and around 84 min in test

#4. Also in test #4, indoor temperature oscillated around the virtual building temperature setpoint possibly due to test facility control which also contributes to observed differences in performance. Similar behavior of indoor temperature control was also observed in some other test intervals at PG&E. It should be further looked into to decide whether it's acceptable to ensure load-based testing repeatability and reproducibility and update test operating tolerances as required. Overall, similar to cooling dry coil results, observed variability in measured performance for 54°F and 47°F continental intervals where unit cycles on/off could be due to variation in test unit dynamic response, thermostat sensing dynamics, airflow rate, test setup and measurements, test facility control along with dynamic test and measurement uncertainties.

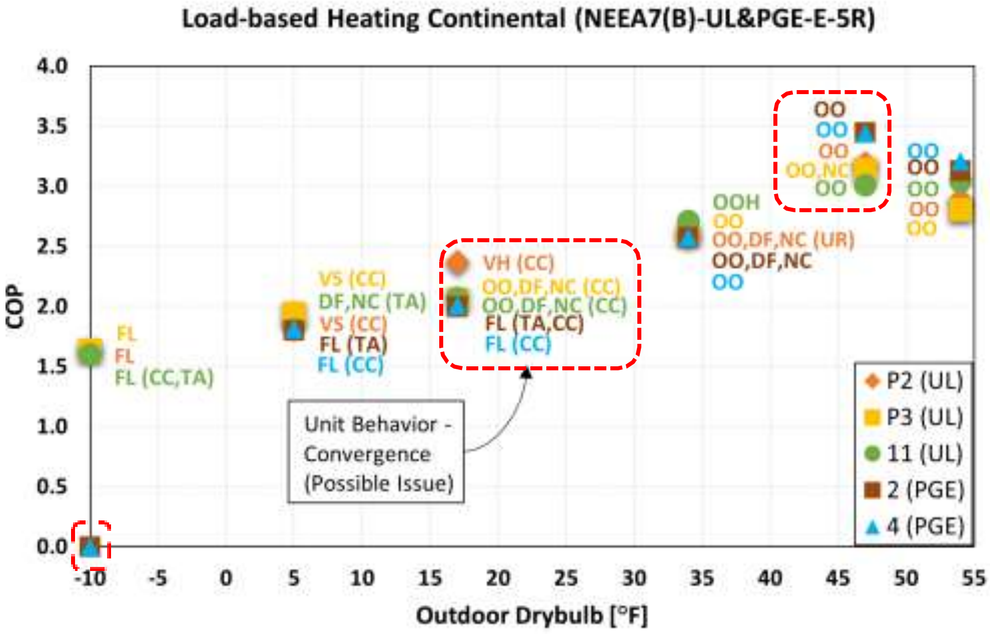


Figure 149. NEEA7(B) heating continental test intervals COP with unit behavior during convergence

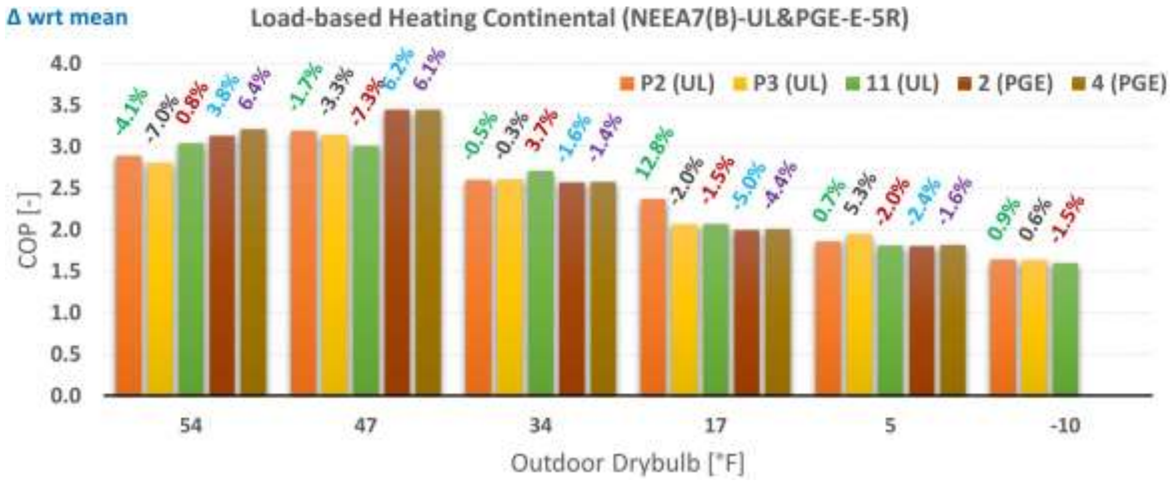


Figure 150. NEEA7(B) heating continental test intervals COP comparison

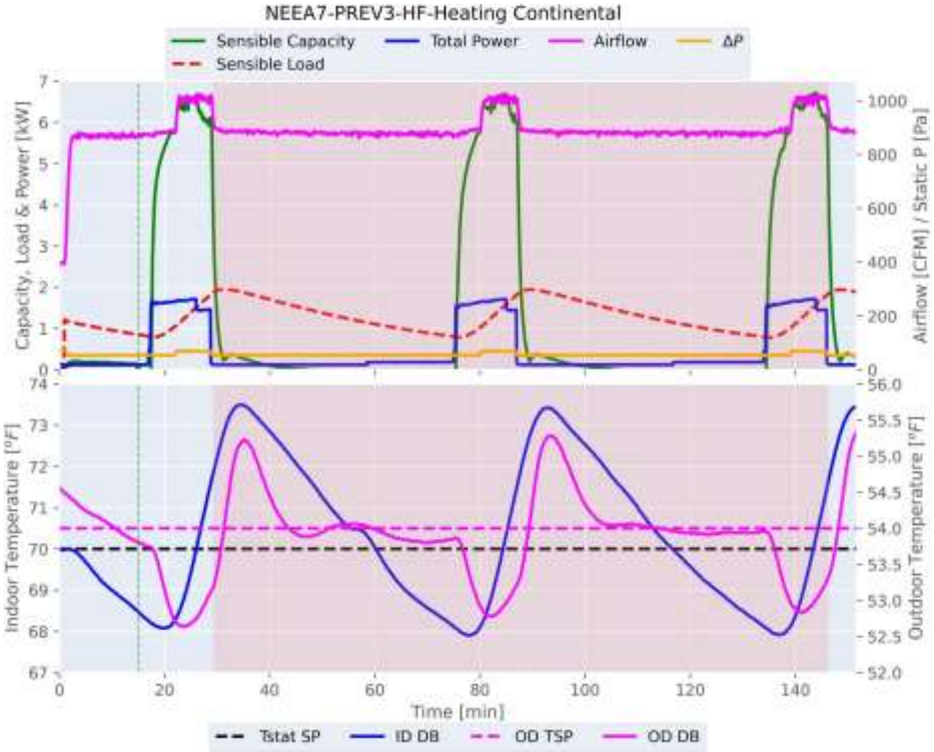


Figure 151. NEEA7-P3 (UL) performance and temperature variations for heating continental test interval HF (54℉)

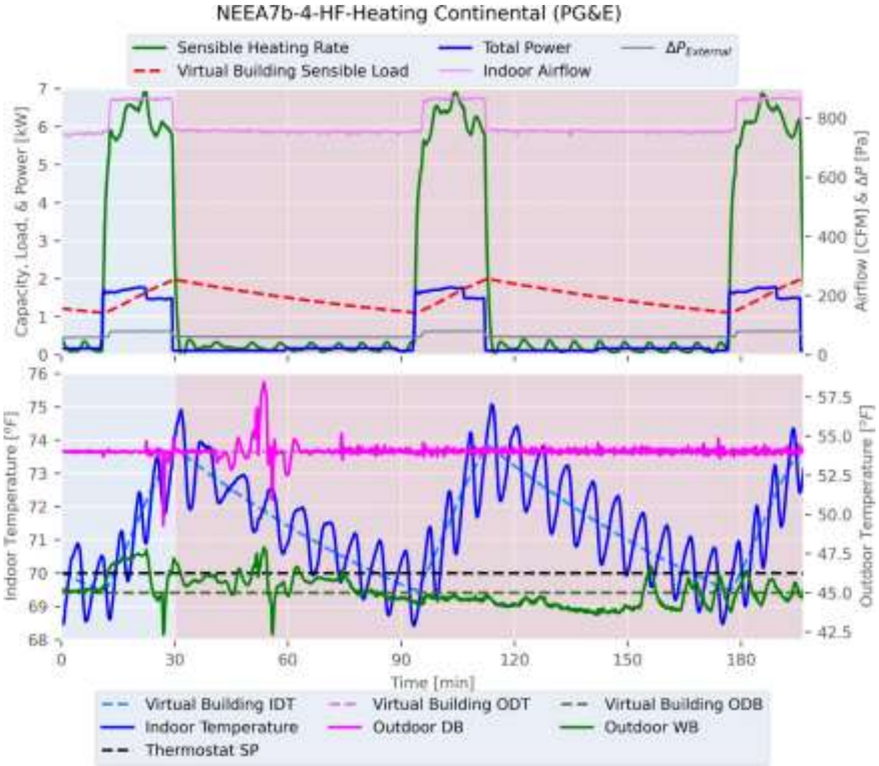


Figure 152. NEEA7B-4 (PG&E) performance and temperature variations for heating continental test interval HF (54℉)

In the 34°F ambient temperature continental test interval, UUT was cycling on/off in all 5 tests, and an intermittent defrost was also observed in tests P2 and #4 where convergence was not found within 6 hours of test duration. Similar to 54°F and 47°F test intervals, between the two labs, differences were observed in test unit dynamic response, indoor temperature variation based on the virtual building model response, cycling rate, airflow rate, and test facility control to maintain test conditions as shown in Figure 153 and Figure 154 for test P3 at UL and #4 at PG&E for HD test interval. The highest COP was observed in test #11 at UL which is possibly due to a relatively higher airflow rate compared to the other 4 tests as shown in Figure 155. This could be due to variation test unit dynamic response with its integrated controller or airflow settings during test unit setup. Nonetheless, irrespective of all the differences outlined above, overall good reproducibility was observed in this test interval. This could be due to some of these differences might be counteracting each other, resulting in comparable performance between different tests unlike 54°F and 47°F test intervals.

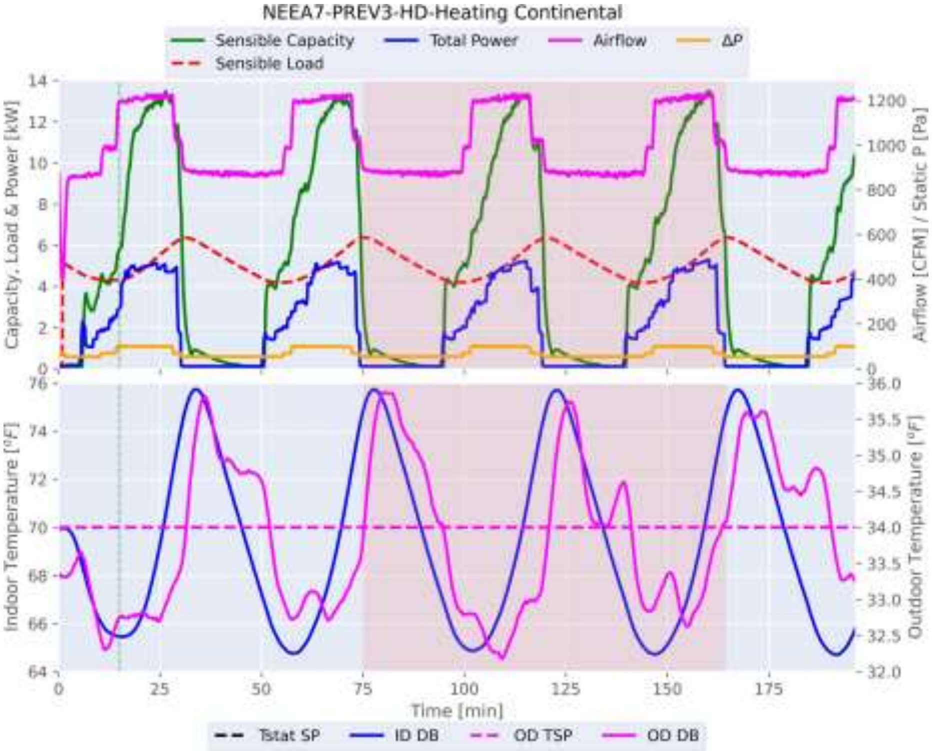


Figure 153. NEEA7-P3 (UL) performance and temperature variations for heating continental test interval HD (34℉)

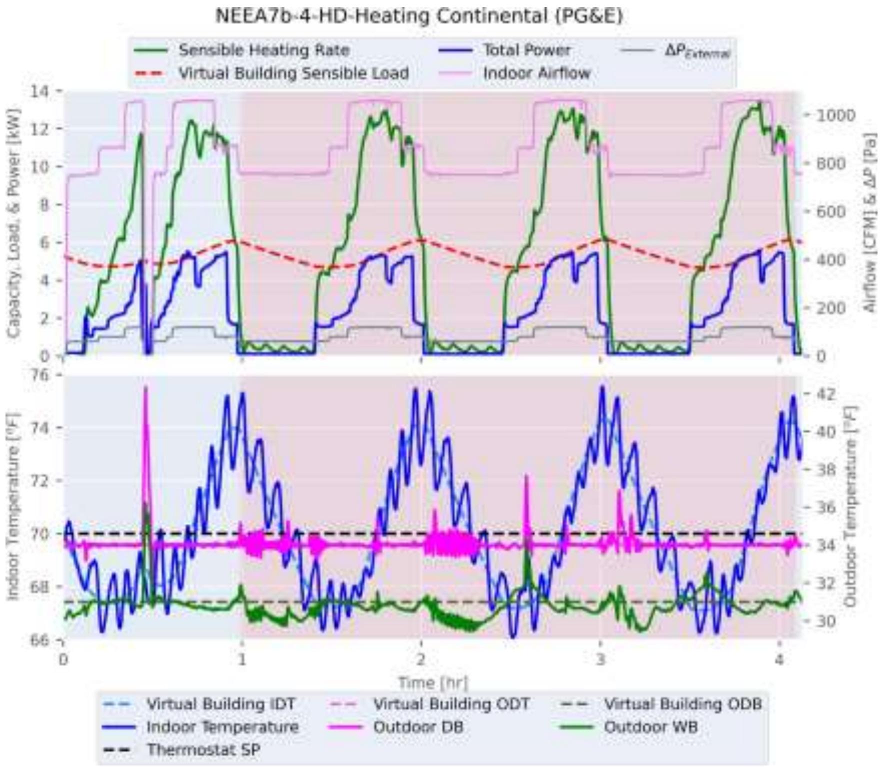


Figure 154. NEEA7B-4 (PG&E) performance and temperature variations for heating continental test interval HD (34℉)

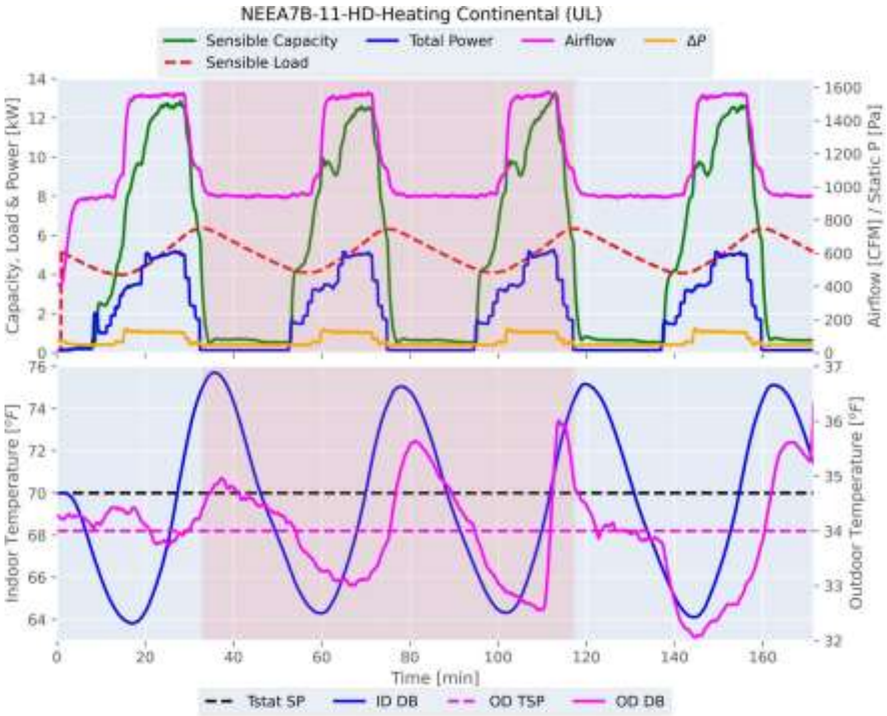


Figure 155. NEEA7B-11 (UL) performance and temperature variations for heating continental test interval HD (34℉)

In the 17°F ambient temperature continental test interval, the test unit showed overall good reproducibility except for higher COP in test P2 at UL compared to the other 4 tests. In this test interval, load-based tests were performed at UL, while full-load tests were conducted at PG&E, and different operation modes were captured during the convergence period in different tests. In UL tests, UUT was cycling on/off with intermittent defrosts as shown in Figure 156 and Figure 157 for tests P2 and #11. In test P2, convergence was found in a 40 min window before the test unit cycled off thus not capturing the on/off cycle and defrost in average performance which is a possible convergence criteria implementation issue. Due to this, higher COP was measured in test P2 compared to test P3 and #11 at UL, where convergence was not found, and complete on/off cycles including defrosts in 6-hour test duration were taken in converged duration to calculate the average performance. Further, in full-load test intervals in tests #2 and #4 at PG&E, UUT was going into defrost, however, convergence was found during steady operation mode in between defrosts as shown in Figure 158 for test #4. Without including defrosts in average performance resulted in relatively better COP in PG&E tests and thus improved reproducibility between two labs due to this possible convergence criteria issue. In between different tests, COP from its mean of 5 tests varied from -5% to 12.8%, which is possibly due to differences in test unit dynamic response, thermostat dynamics, indoor temperature variation, testing approach (load-based vs full-load), airflow rate, test setup, and instrumentation. In continental test interval at 5°F outdoor temperature, relatively good reproducibility was noted with differences in COP from its mean of 5 tests varying from -2.4% to 5.3%. However, it should be noted the during convergence period defrost was only captured in test #11 due to possible convergence criteria implementation issues similar to the 17°F test interval, which might impact the actual reproducibility if defrost was included in all test's average performance. Further, at -10°F ambient temperature, test unit performance was only captured at UL and not at PG&E due to test facility limitations.

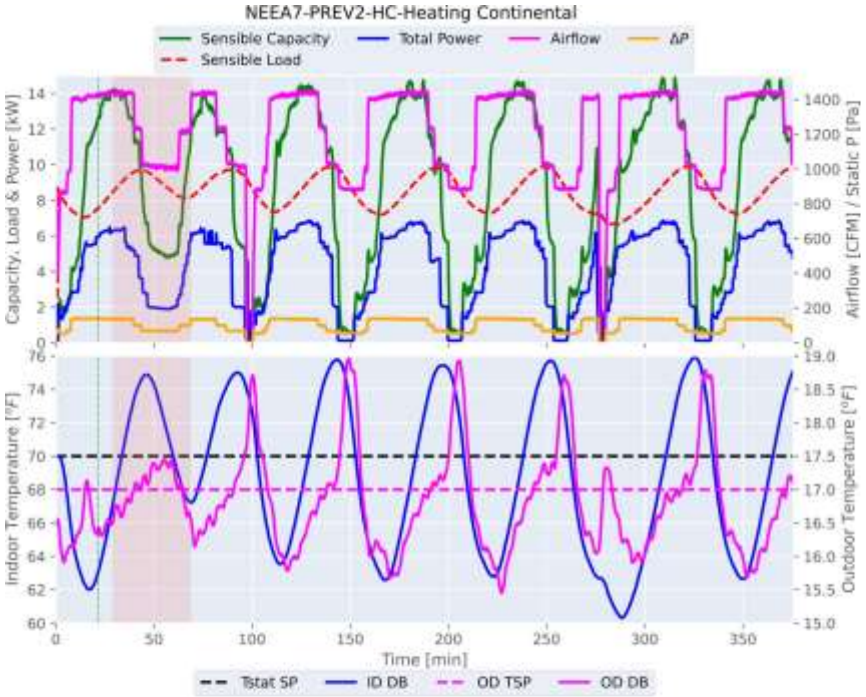


Figure 156. NEEA7-P2 (UL) performance and temperature variations for heating continental test interval HC (17°F)

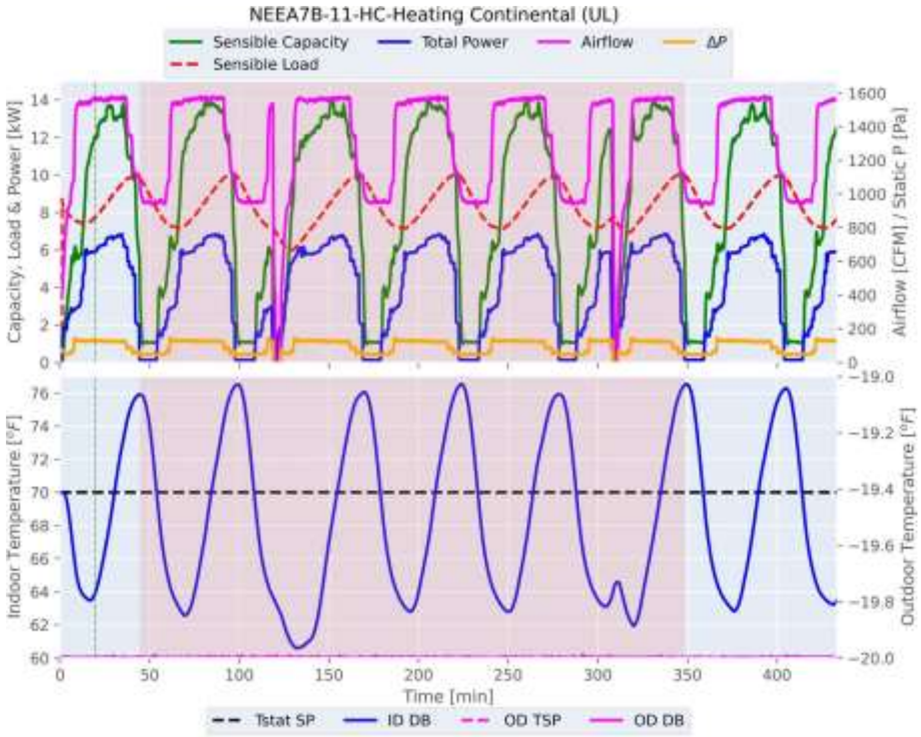


Figure 157. NEEA7B-11 (UL) performance and temperature variations for heating continental test interval HC (17℉)

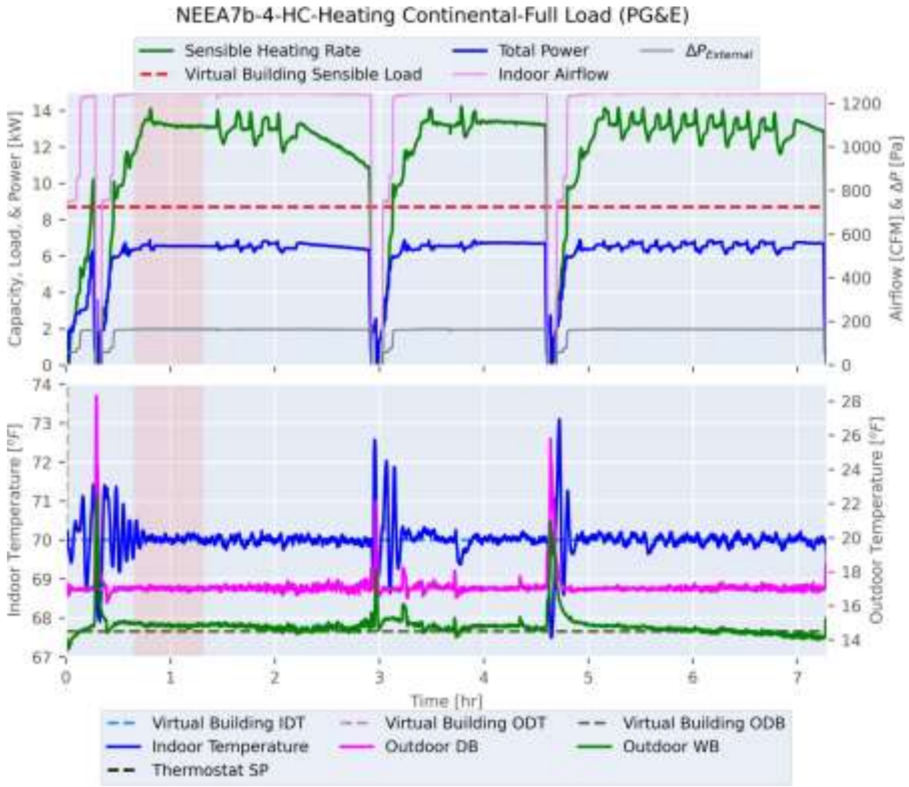


Figure 158. NEEA7B-4 (PG&E) performance and temperature variations for heating continental test interval HC (17℉)

Figure 159 and Figure 160 show the comparison of NEEA7(B) performance in heating marine test intervals for five sets of tests. Similar to continental test intervals, the test unit cycled on/off at 54°F and 47°F outdoor temperatures, and a larger variation in COP was observed with relatively better performance observed in tests at PG&E. In the 34°F ambient temperature marine test interval, relatively good reproducibility was observed; however, in tests at UL, an intermittent defrost was observed in between on/off cycles, and the test unit just cycled on/off without any defrost. Also, in the majority of the tests at 34°F outdoor temperature, no convergence was achieved in the 6-hour maximum test duration.

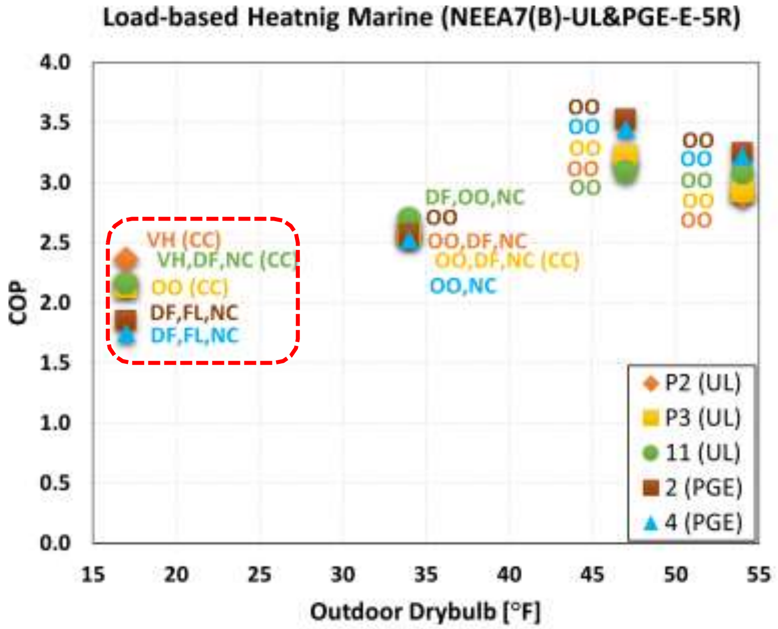


Figure 159. NEEA7(B) heating marine test intervals COP with unit behavior during convergence

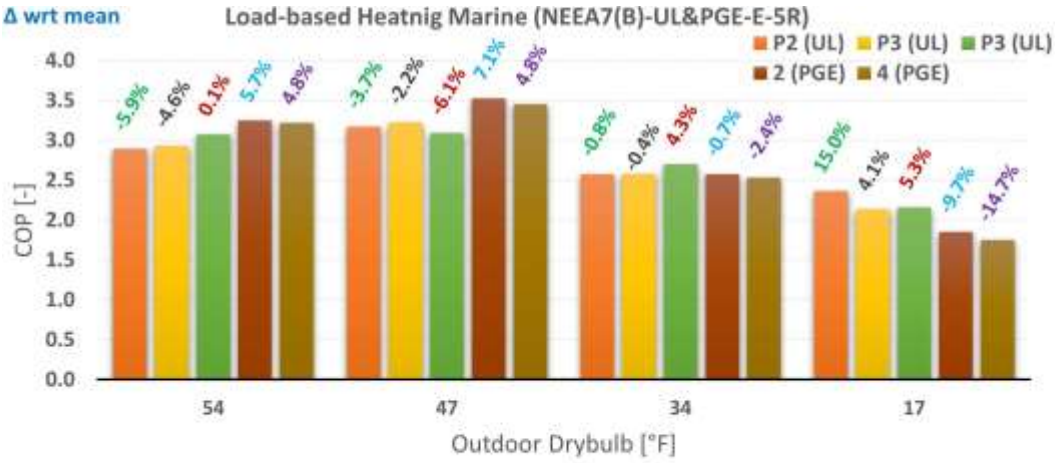


Figure 160. NEEA7(B) heating marine test intervals COP comparison

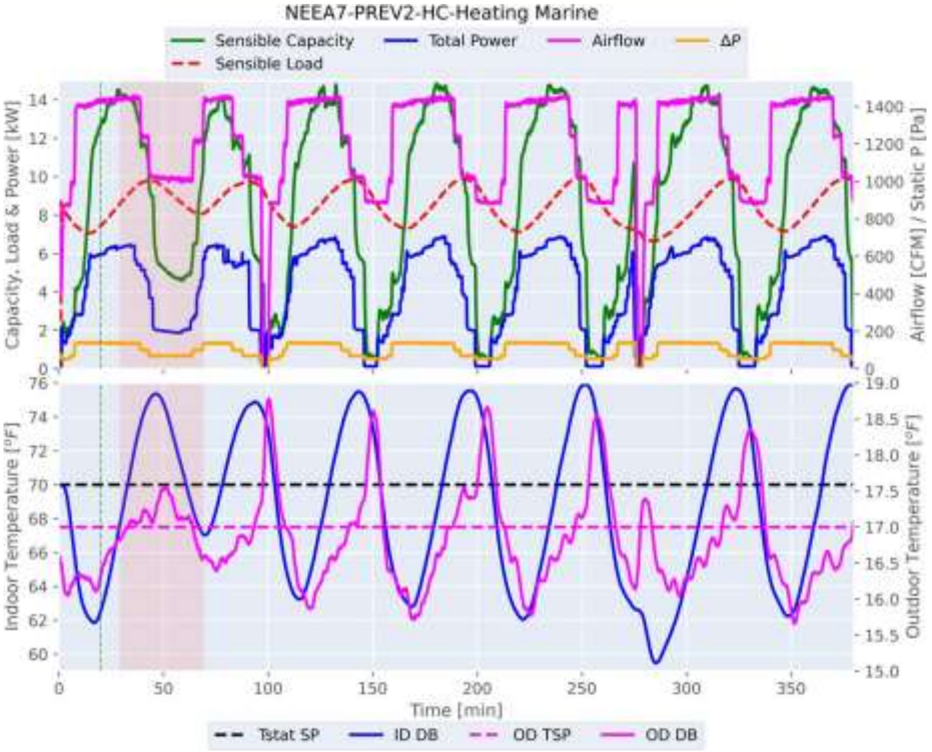


Figure 161. NEEA7-P2 (UL) performance and temperature variations for heating marine test interval HC (17℉)

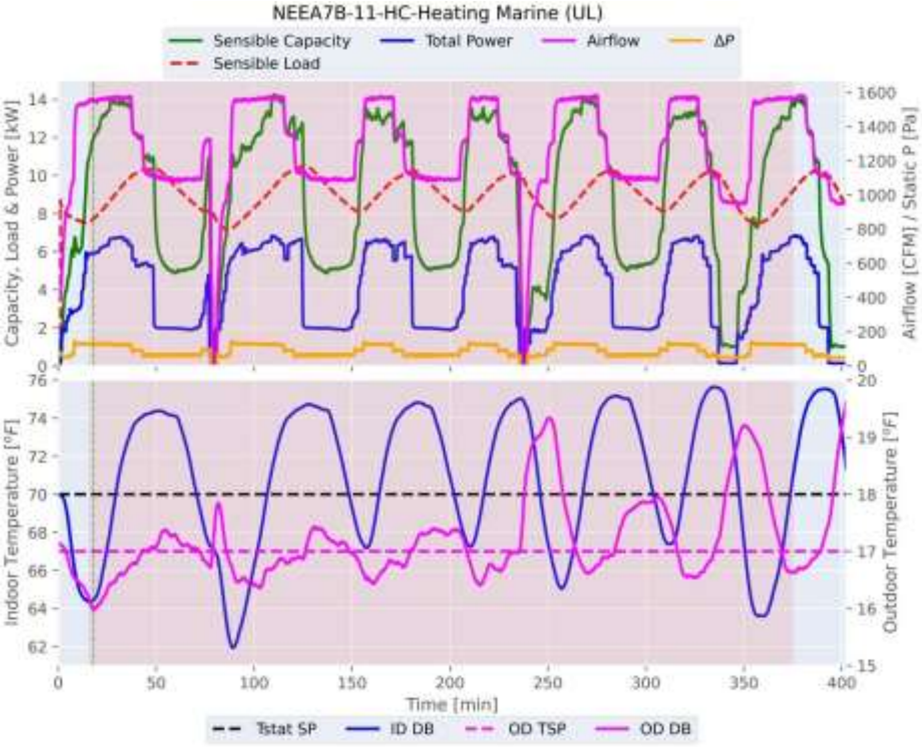


Figure 162. NEEA7B-11 (UL) performance and temperature variations for heating marine test interval HC (17℉)

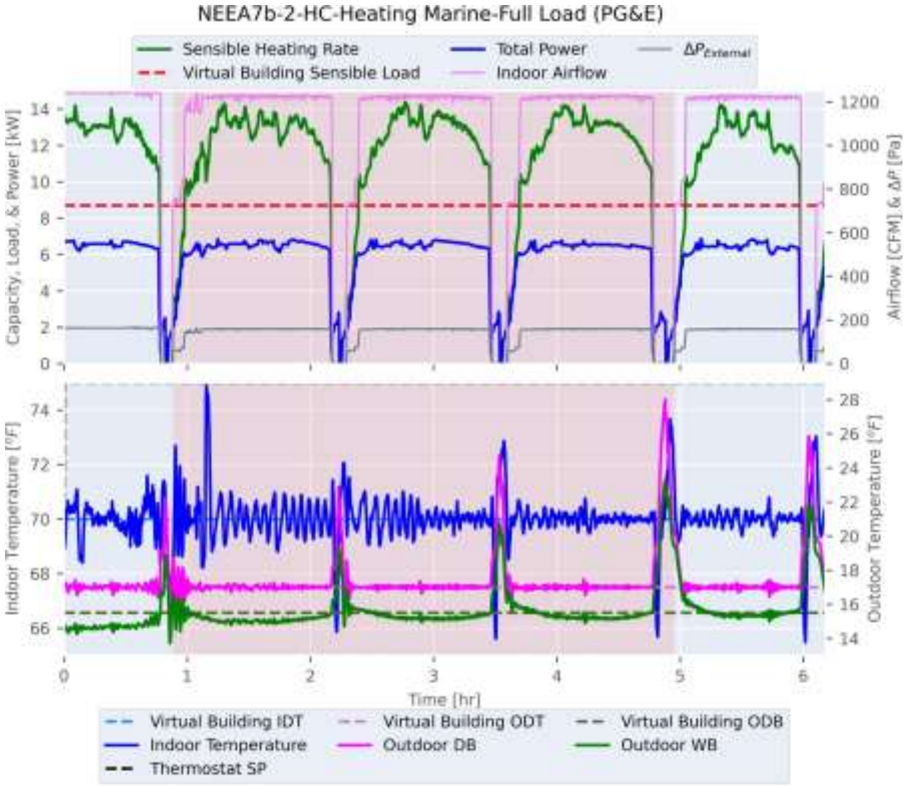


Figure 163. NEEA7B-2 (PG&E) performance and temperature variations for marine continental test interval HC (17°F)

In the 17°F ambient temperature marine test interval, poor reproducibility was noted with the test unit performing better in load-based tests at UL compared to full-load tests at PG&E under the same test conditions. Comparatively better COP in test P2 at UL is mainly due to convergence criteria not capturing on/off cycles and defrosts in average performance as shown in Figure 161 which shows possible convergence criteria issues similar to continental test interval. Figure 162 and Figure 163 show the test unit performance at the same test conditions in test #11 at UL and test #2 at PG&E. As can be seen that similar to continental test intervals, differences in test unit dynamic response, indoor temperature variation, and airflow rate were also observed in marine test intervals. Further at 17°F ambient temperature relatively large difference in COP was observed between test #11 and test #2 compared to continental test interval under the same conditions, because here in test #2 converged period accounted for defrost in an average performance, unlike continental test interval. This resulted in lower performance in tests #2 and #4 at PG&E and a larger difference compared to UL tests. Overall similar factors as discussed in continental test intervals played the part in observed differences for marine test intervals.

Figure 164 shows the average COP based on 5 sets of tests in both labs for continental and marine test intervals along with STD and the maximum COP differences for each test interval as a percentage of the mean value. For continental test intervals, the maximum difference in COP varies from 2.5% to 17.7% and STD from ±1.1% to ±6.5%; whereas for marine test intervals maximum difference in COP varies from 6.6% to 29.7% and STD from ±2.2% to ±10.8%. Also, in continental and marine test intervals similar average COPs were noted at the same outdoor conditions, showing that the higher outdoor humidity in marine test intervals does not affect this test unit's performance significantly. Also, it should be noted that test unit performance was not measured in PG&E at -10°F outdoor temperature, so here the results for that test interval basically show test unit repeatability at UL.

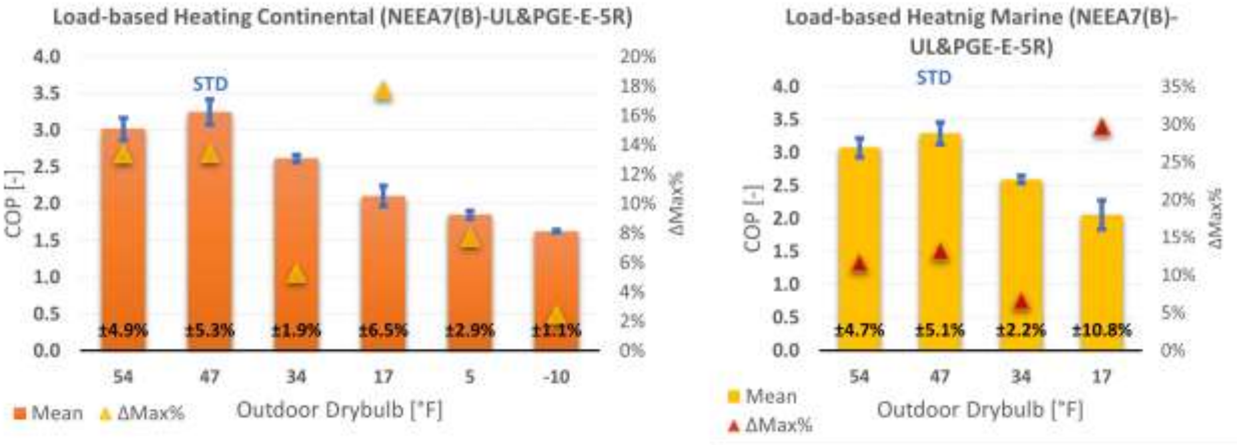


Figure 164. NEEA7(B) heating continental and marine test interval COP average and maximum difference for different tests with a standard deviation

Figure 165 and Figure 166 show the comparison of estimated heating SCOP for different climate zones based on the test results of five tests. Overall good reproducibility was observed in heating SCOP across different climate zones with maximum differences in SCOP between different tests varying from 3% to 4.8%, 95% CI from ±1.7% to ±2.5%, and STD from ±1.2% to ±1.8%. In general, for the "Marine" climate zone which utilizes marine test intervals results in SCOP estimation, the highest SCOP was based on test #2 results at PG&E and the lowest based on test P2 and UL. Whereas for other climate zones which utilize continental test intervals results, the highest SCOP was estimated based on test P2 results at UL and the lowest based on test #11 results at UL.

Table 21 shows the overall SCOP reproducibility results summary for NEEA7(B)-UL&PGE-E-5R dataset with the range and mean of statistical parameters. Overall test unit showed good reproducibility in heating mode test results; however poor reproducibility in cooling mode test results, especially for climate zones that utilize cooling dry coil test intervals results in SCOP estimation.

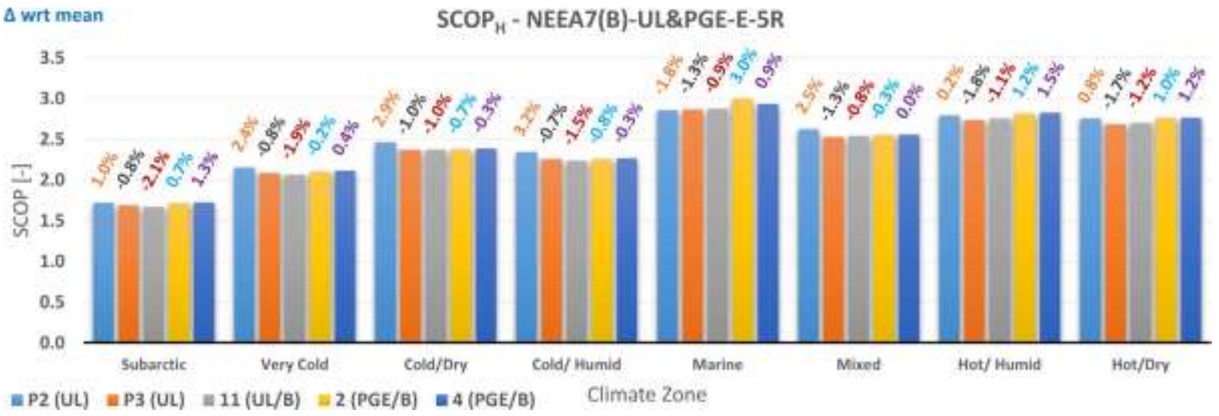


Figure 165. NEEA7(B) heating SCOP comparison in different climate zones for different tests

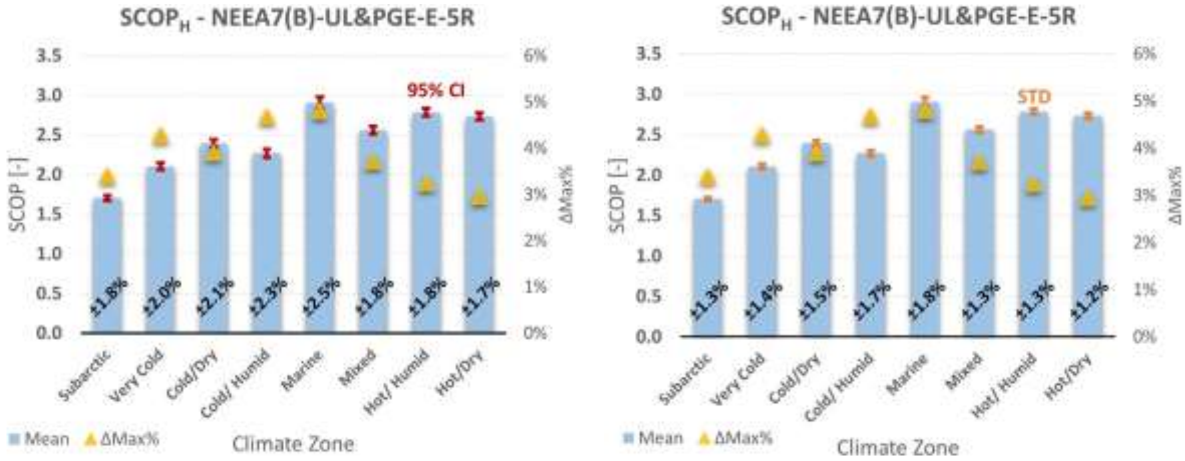


Figure 166. NEEA7(B) heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

Table 21. NEEA7(B)-UL&PGE-E-5R dataset SCOP reproducibility summary results

Metric	**No. of Tests**	**∆SCOP (Max-Min) [%]**			**SCOP 95% CI [%]**			**SCOP STD [%]**		
		Min	***Max***	***Mean***	***Min***	***Max***	***Mean***	***Min***	***Max***	***Mean***
$SCOP_C$	5	8.0%	24.1%	14.6%	4.2%	13.7%	8.1%	3.0%	9.8%	5.8%
$SCOP_H$	5	3.0%	4.8%	3.9%	1.7%	2.5%	2.0%	1.2%	1.8%	1.4%

5.2 NEEA4

In this section, the EXP07 reproducibility assessment is presented for the NEEA4 unit, a 1-ton mini-split ductless variable-speed heat pump, based on five sets of test results at UL and PG&E lab in NEEA4-UL&PGE-E-5R dataset. Three tests (P1, #8, and #10) were performed at UL, and two (#6 and #8) were performed at PG&E. First, test unit performance was measured in test P1 at UL around January – February 2019. Then unit was taken out of the test rooms and re-installed to perform two back-to-back tests, #8 and #10, at UL around June – July 2020. Then test unit was shipped to PG&E and installed in the lab there to perform other two back-to-back tests, #2 and #4, around June – August 2021. In the following subsections, first, cooling mode results are presented followed by heating mode reproducibility assessment results for NEEA4.

5.2.1 Cooling Mode

NEEA4 performance comparison for cooling dry coil test intervals is shown in Figure 167 and Figure 168 along with test unit operation mode during convergence and any possible issue with test interval. In general, relatively lower COPs were measured at PG&E compared to UL tests, similar to NEEA(B) cooling dry coil intervals test results. In 77°F outdoor temperature test interval (CE), the test unit was cycling on/off in all tests, and COP from the mean of 5 tests varied from -5.2% to 7.4%. At 86°F outdoor temperature test interval (CD), the test unit was also cycling on/off cycling in all five tests, however, test #8 at PG&E convergence was found in variable speed mode before the test unit cycled off as shown in Figure 169, thus not capturing the on/off cycles in average performance which shows a possible issue with convergence criteria implementation. In test interval CD, COP from the mean of 5 tests varied from -5.2% to 5.2% across different tests, similar to CE test interval. This variation in test unit performance in these two intervals could be due to differences in test unit dynamic response, thermostat sensing dynamics, test setup including refrigerant charge and instrumentation. For example, Figure 169, Figure 170, and Figure 171 show the test unit performance at 86°F outdoor temperature in test #8 at PG&E, test P1 at UL, and test #10 at UL, respectively. At PG&E, relatively lower airflow and sensible cooling rate were measured during compressor *on* period compared to UL results. This difference in measured airflow is similar to what was noted in the NEEA7(B) test results. Even in two tests at UL, P1 and #10, there is a difference in airflow, which could be because of the test setup as the unit was re-installed after test P1. Also, a difference in indoor temperature average, as well as its range of change during on/off cycling, can be seen in three tests. This is due to both variation in thermostat sensing dynamics as well test unit dynamic response which affect the sensible cooling rate and thus indoor temperature change based on virtual building model response. The effect of all these differences could be seen in test unit cycling rate response in three tests.

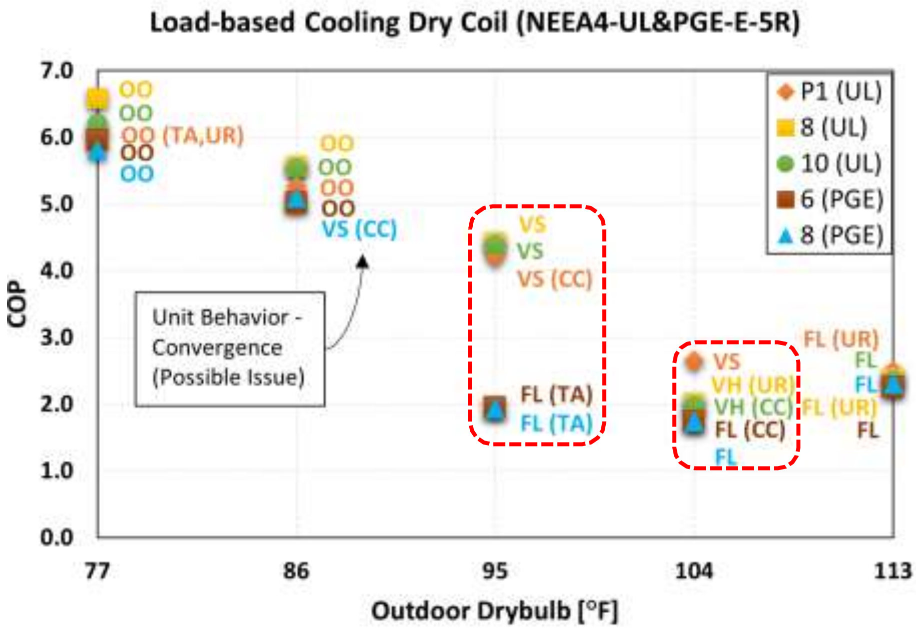


Figure 167. NEEA4 cooling dry coil test intervals COP with unit behavior during convergence

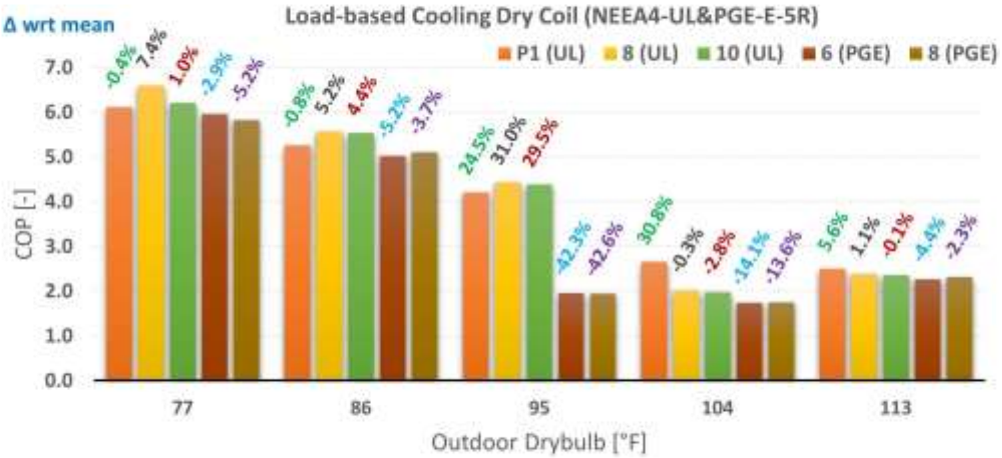


Figure 168. NEEA4 cooling dry coil test intervals COP comparison

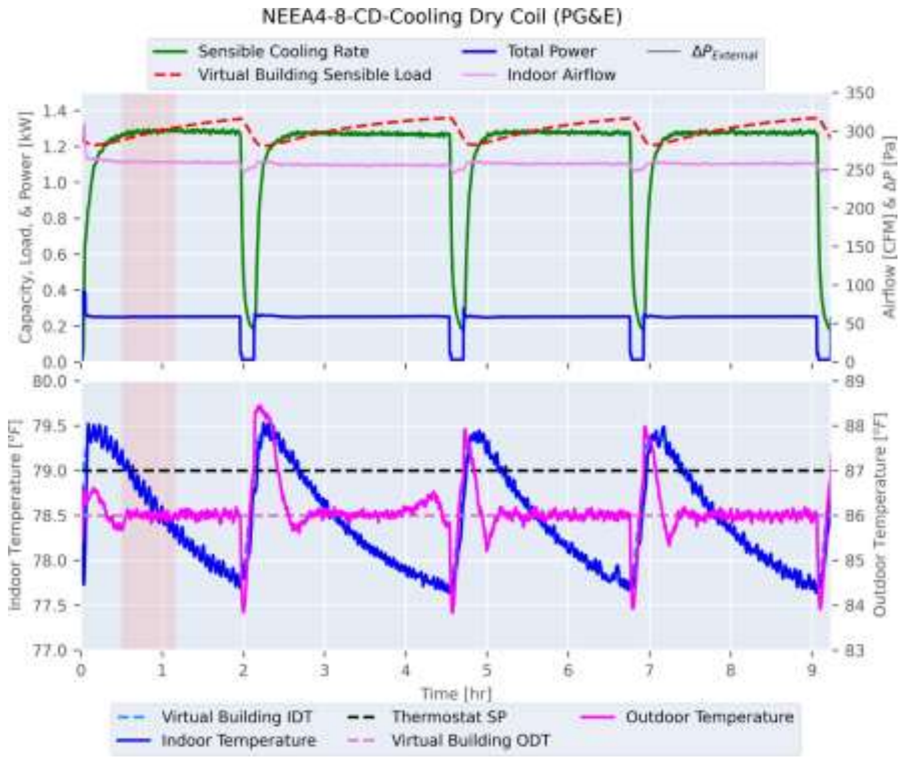


Figure 169. NEEA4-8 (PG&E) performance and temperature variations for cooling dry coil test interval CD (86℉)

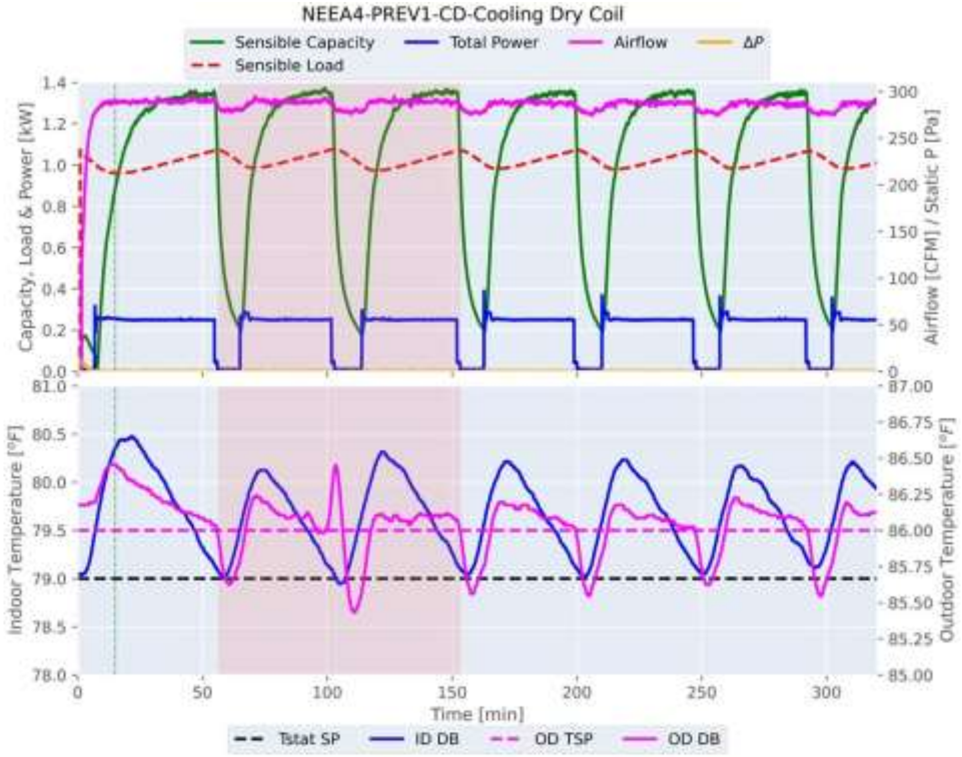


Figure 170. NEEA4-P1 (UL) performance and temperature variations for cooling dry coil test interval CD (86℉)

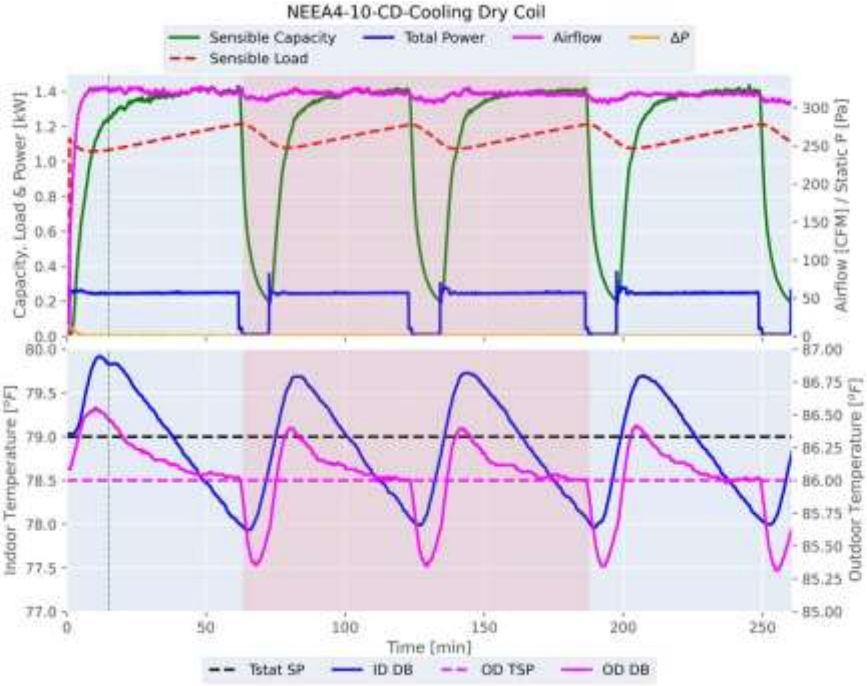


Figure 171. NEEA4-10 (UL) performance and temperature variations for cooling dry coil test interval CD (86°F)

In the cooling dry coil test interval at 95°F ambient temperature, there is a large difference in performance between the two test facilities with relatively higher COP in UL tests compared to PG&E tests. This difference is mainly because of the difference in testing approach at this test condition between the two labs. At UL load-based tests were performed, and the test unit was operating in variable speed mode where convergence was found in a 40 min window as shown in Figure 172 for test #8 at UL. However, at PG&E, full-load tests were performed by deactivating the virtual building model and making the unit run at maximum capacity as shown in Figure 173 for test #6 at PG&E. This resulted in around 73% lower measured COP in test #6 at PG&E compared to test #8 at UL. In test #6 at PG&E, due to slow compressor ramp-up of the unit at the beginning of the test resulted in indoor temperature higher than the full-load test limit of 81°F for dry coil tests. This led to the decision to change to a full-load test even though the unit was not running out of capacity as can be seen by comparing the sensible cooling rate and building load later in the test. This shows a possible issue with the clarity of the current testing approach as per EXP07 for cases like this. Nonetheless, because of this difference in testing approach between the two labs at this test condition, these results could not be used for reproducibility assessment of the load-based testing approach.

A similar difference in testing approach between the two labs could be seen at 104°F outdoor temperature, however, relatively smaller differences in performance were observed due to higher virtual building load. Further in UL tests, higher COP was measured in test P1 compared to tests #8 and #10 due to the difference in test unit dynamic response in variable speed mode. Further, variation in test unit measured performance in full-load tests at 113°F outdoor temperature shows the differences due to test unit setup, instrumentation, airflow measurement, etc. without the effect of test unit dynamics response with its thermostat and embedded controller. In this interval, higher performance was measured at UL with the difference in COP from the mean of 5 tests varying from -4.4% to 5.6%.

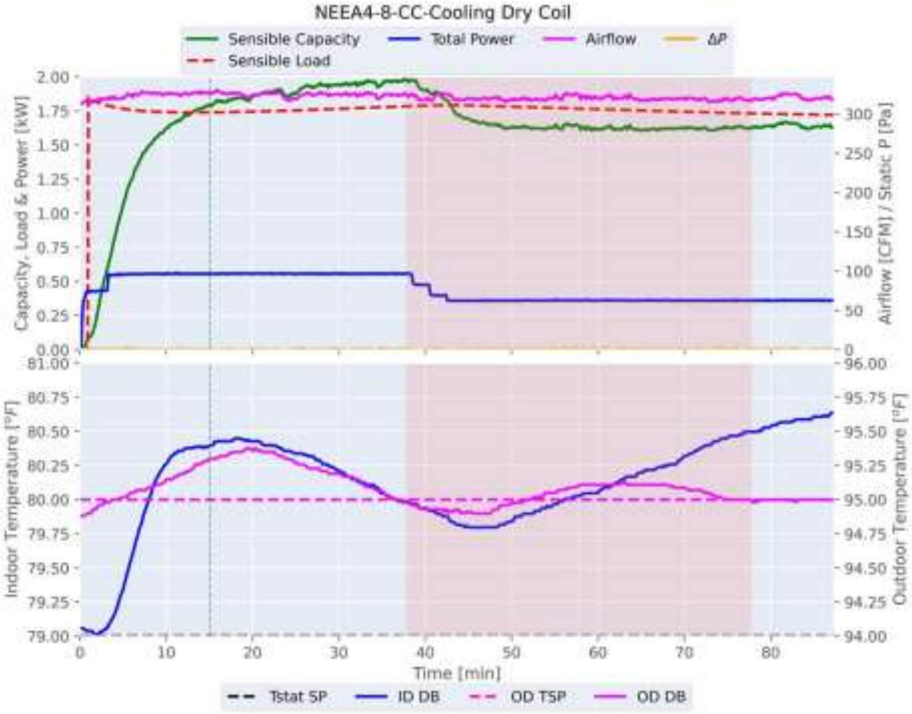


Figure 172. NEEA4-8 (UL) performance and temperature variations for cooling dry coil test interval CC (95°F)

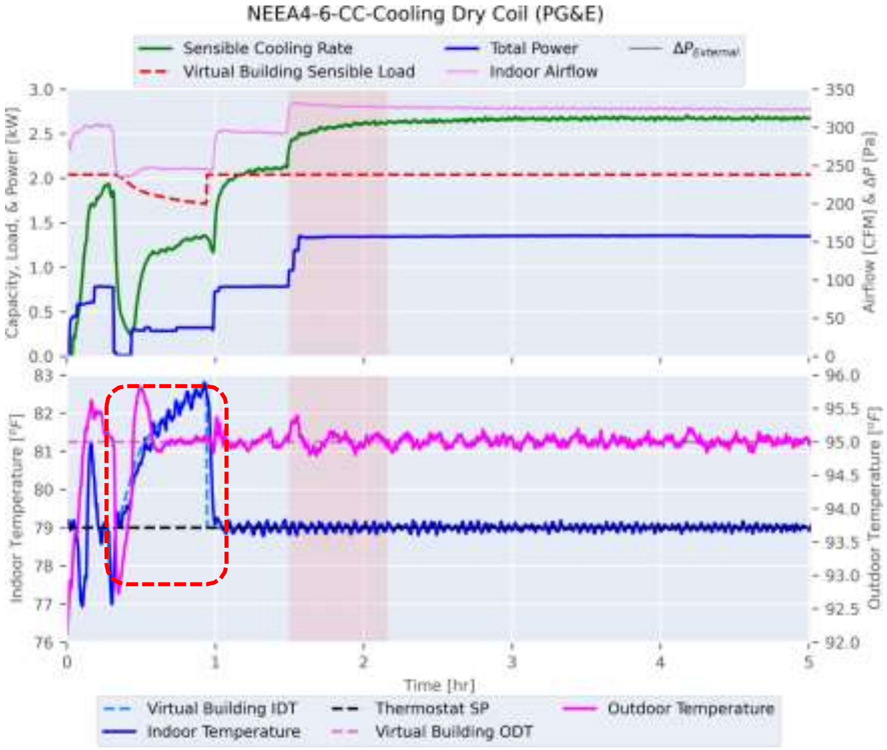


Figure 173. NEEA4-6 (PG&E) performance and temperature variations for cooling dry coil test interval CC (95°F)

Figure 174 and Figure 175 show NEEA4 performance comparison among 5 tests for cooling humid coil intervals. In 77°F outdoor temperature test interval, the test unit cycled on/off in all 5 tests, and relatively better COPs were measured in PG&E tests, in contrast to cooling dry coil test interval results. The differences in COP from the mean of 5 tests varied from -2.4% to 4%. In addition to the observed differences in test unit sensible cooling rate, airflow, indoor temperature between two labs as seen in dry coil test intervals at similar test conditions, here in humid coil tests, a difference in indoor humidity was also observed in two labs which resulted in higher latent cooling rate in PG&E tests and thus possibly better performance. Figure 176 and Figure 177 show the test unit performance with time for test #8 UL and test #6 at PG&E, respectively. Here convergence criteria issue is related to the average performance period not capturing complete on/off cycles in a 4-hour test duration in case of non-convergence.

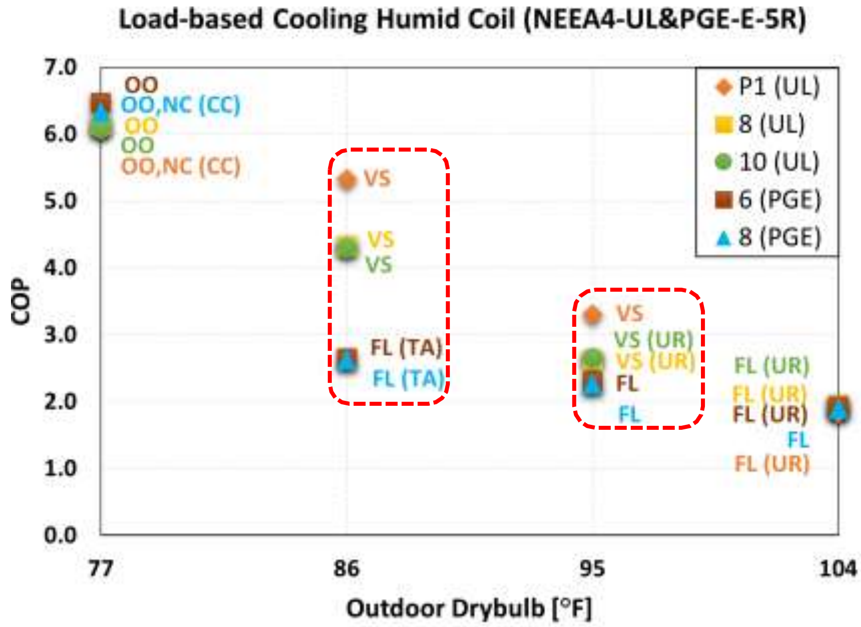


Figure 174. NEEA4 cooling humid coil test intervals COP with unit behavior during convergence

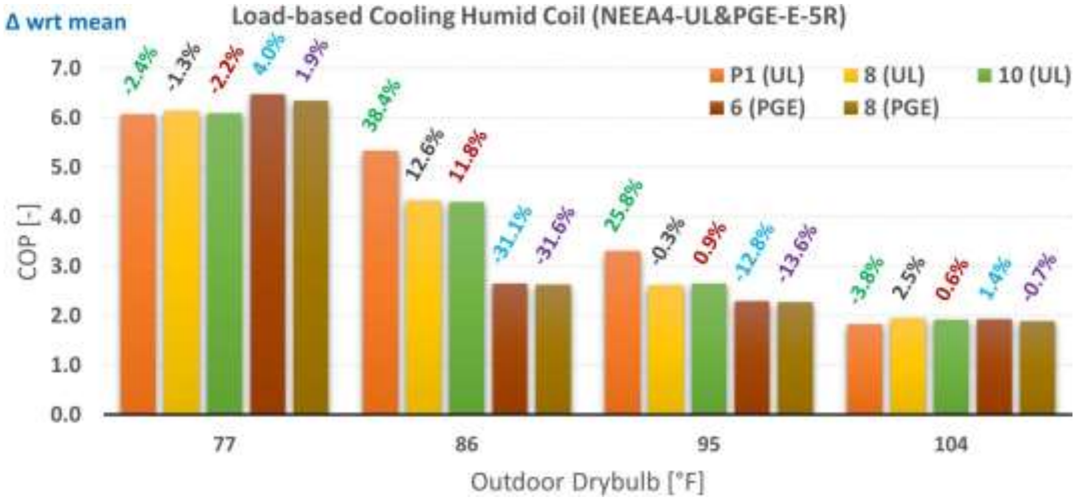


Figure 175. NEEA4 cooling humid coil test intervals COP comparison

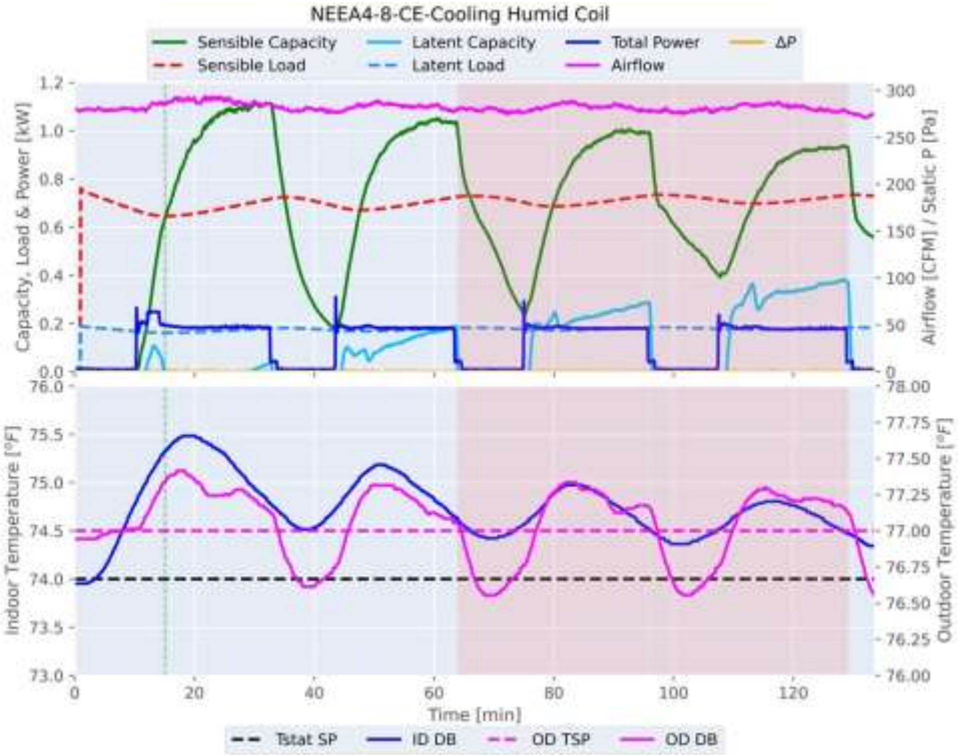


Figure 176. NEEA4-8 (UL) performance and temperature variations for cooling humid coil test interval CE (77°F)

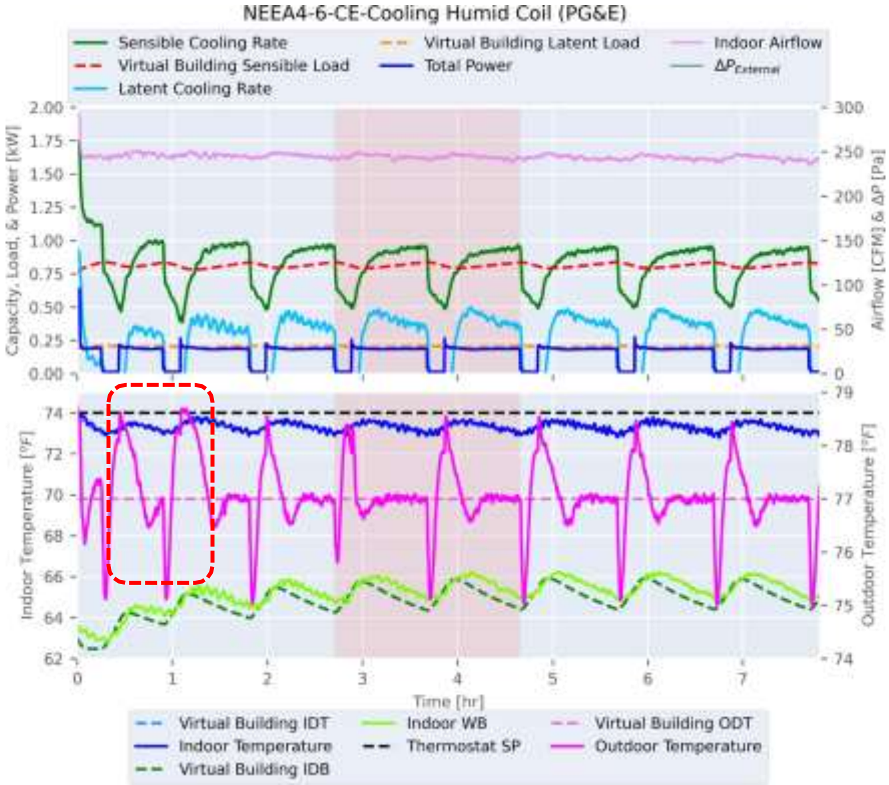


Figure 177. NEEA4-6 (PG&E) performance and temperature variations for cooling humid coil test interval CE (77°F)

In 86°F and 95°F outdoor temperature humid coil test intervals, a large variation in performance was observed between the two labs mainly due to difference in testing approach, a similar issue as seen in 95°F and 104°F cooling dry coil test intervals, where load-based tests were performed at UL with the unit operating in variable speed mode however full-load test conducted at PG&E resulting in relatively lower COP. Further even among UL tests, relatively higher COP was measured in test P1 compared to the other two tests. This was mainly due to differences in test unit dynamic response, indoor temperature variation, and convergence period. In test P1, convergence was found in a period where the test unit was operating at lower compressor speed and higher indoor temperature compared to the other two tests resulting in relatively higher COP. This difference in test unit dynamic response in test P1 compared to tests #8 and #10 at UL could be due to something being changed in test setup during its re-installation to perform tests #8 and #10 after around 16 months after test P1. Further, in full-load tests at 104°F ambient temperature, relatively better reproducibility was observed compared to cooling dry coil full-load tests at 113°F, with the difference in COP from the mean of 5 tests varying from -3.8% to 2.5%. Nevertheless, because of the difference in testing approach in the two labs at 86°F and 95°F outdoor temperature test intervals, these test results could also not be used to perform an overall reproducibility assessment of EXP07 between the two labs.

Figure 178 shows NEEA4 overall COP average of 5 tests for cooling dry coil and humid coil test intervals along with standard deviation and maximum percentage difference in COP. In dry coil test intervals at 77°F, 86°F, and 113°F the maximum difference in COP varies from 10% to 12.5% and STD from ±3.4% to ±4.3%, showing relatively better reproducibility compared to NEEA7(B) at same test conditions. However, larger differences were observed in 95°F and 104°F dry coil test intervals due to the difference in testing approach between the two labs as discussed above. Similarly, large differences were also noted in 86°F and 95°F humid coil intervals due to the same testing approach issue. Because of this, these 4 intervals' results could not be used in reproducibility assessment. Further for 77°F and 104°F humid coil intervals, the maximum difference in COP of 5 tests varies from 6.3% to 6.4% and STD from ±2.2% to ±2.5%, showing relatively better reproducibility compared to dry coil intervals at similar test conditions. Overall, excluding the 4 problematic test intervals, NEEA4 shows better reproducibility compared to NEEA7(B) in dry coil intervals and slightly poor for humid coil intervals

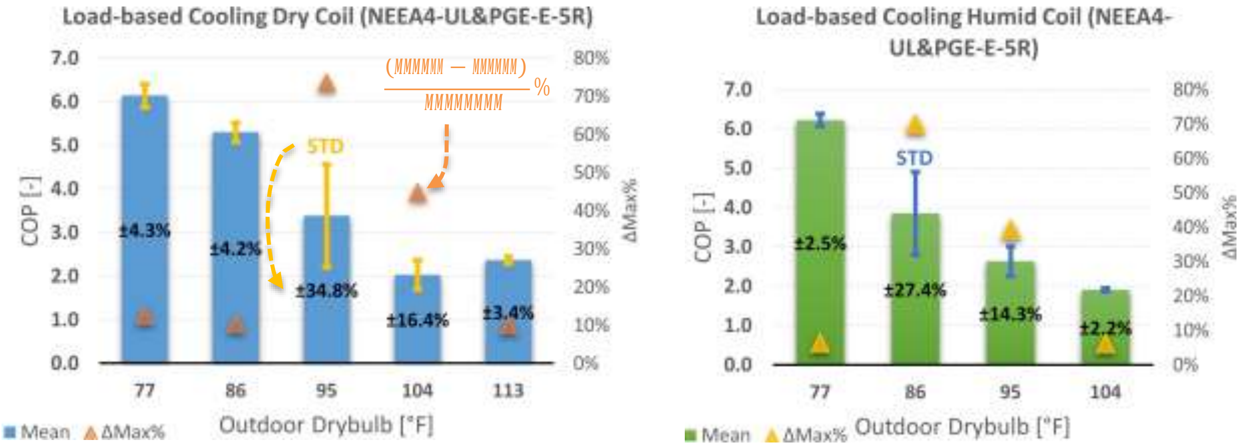


Figure 178. NEEA4 cooling dry coil and humid coil test interval COP average and maximum difference for different tests with a standard deviation

Figure 179 and Figure 180 show the comparison of estimated cooling SCOP based on 5 sets of tests for different climate zones. As it can be seen that there is a large difference in SCOP for test P1 at UL, tests #8 & #10 at UL, and tests #6 and #8 at PG&E. However, because of possible testing approach issues in 4 test intervals as discussed above, these results do not provide any useful conclusions on load-based testing approach reproducibility for NEEA4 and are here only presented for completeness.

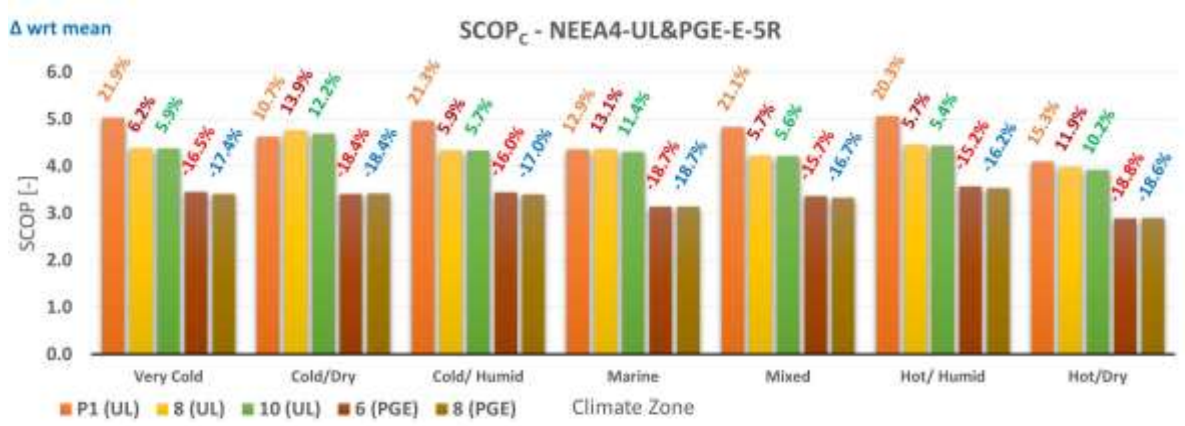


Figure 179. NEEA4 cooling SCOP comparison in different climate zones for different tests

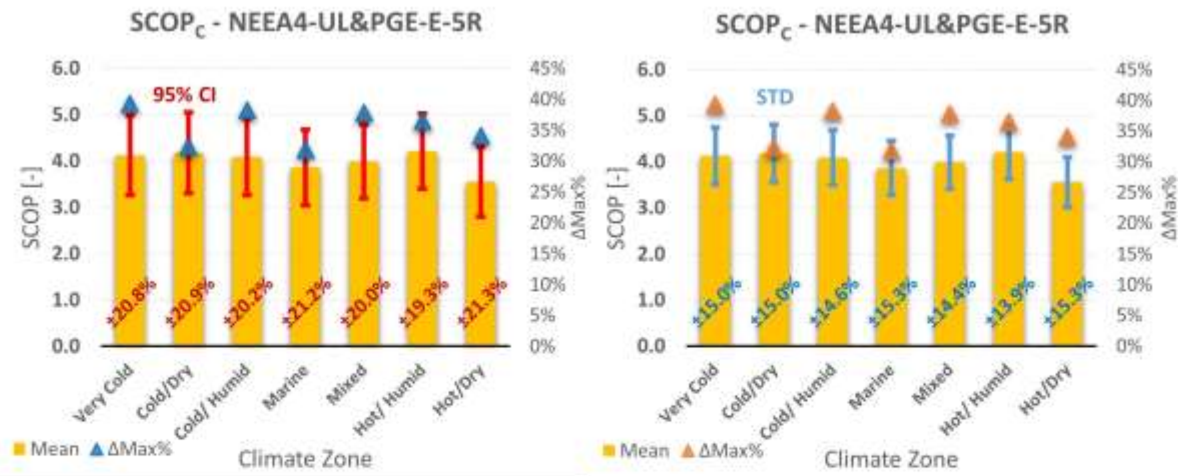


Figure 180. NEEA4 cooling SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

5.2.2 Heating Mode

Figure 181 and Figure 182 show COP comparison of NEEA4 performance and unit operation mode during convergence for 5 tests at UL and PG&E. In quite a few test intervals, possible issues related to convergence criteria, testing approach, and inconsistent unit dynamic response were noted, some of which are discussed in detail in sections 4.2.2 and 4.3.2. At PG&E, similar to NEEA7(B), NEEA4 performance was not measured at the lowest outdoor temperature of -10°F. In general, poor reproducibility was observed in heating continental test intervals with the test unit performing relatively better at PG&E compared to UL in 54°F test intervals where the test unit cycles on/off, whereas performing poorer in low ambient temperatures from 34°F to 5°F.

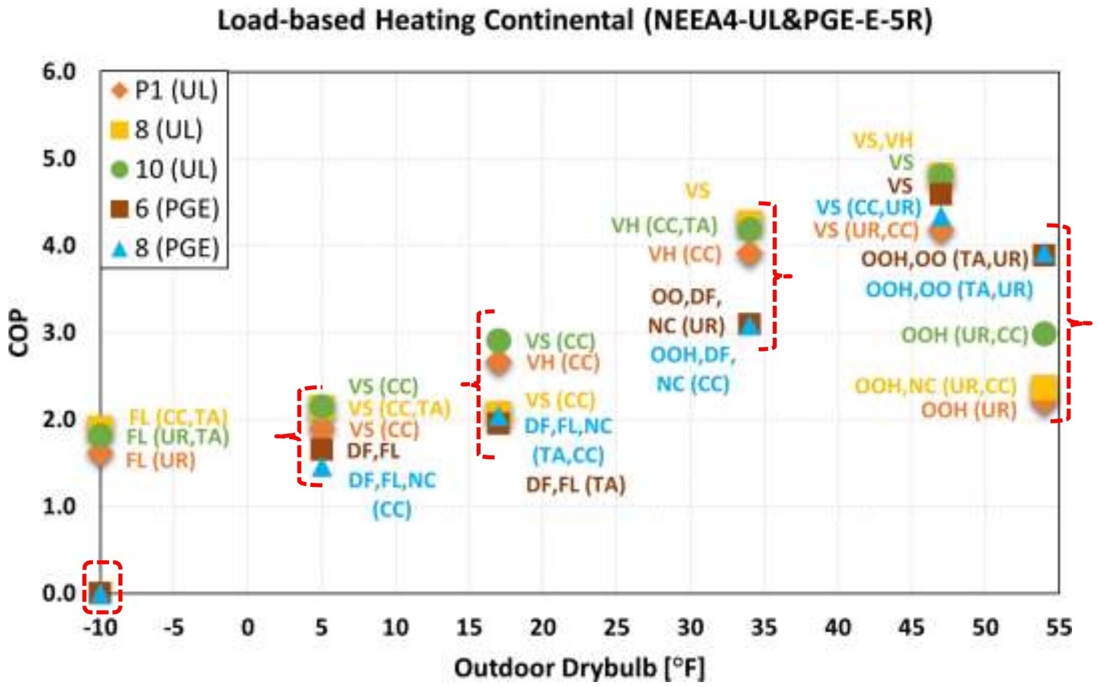


Figure 181. NEEA4 heating continental test intervals COP with unit behavior during convergence

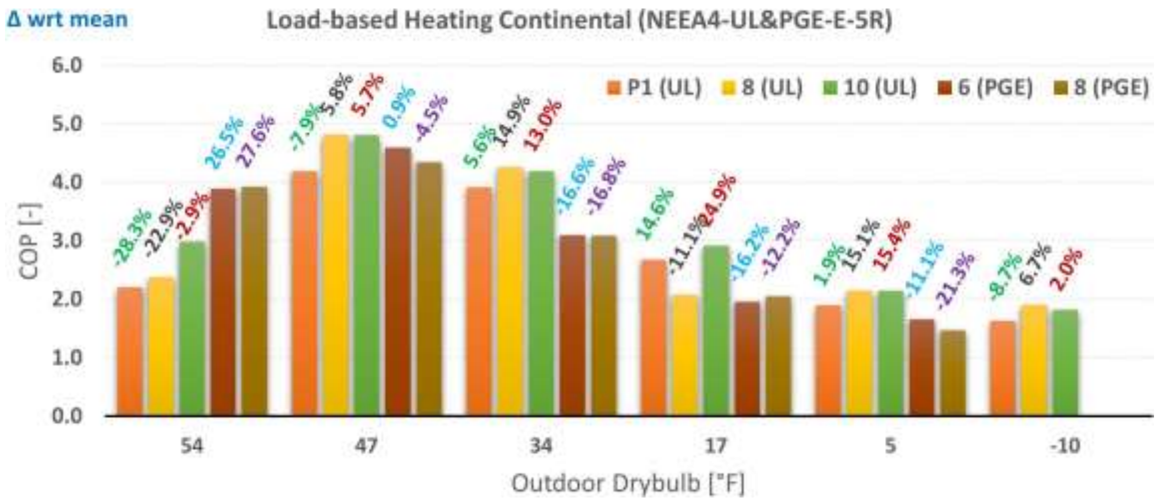


Figure 182. NEEA4 heating continental test intervals COP comparison

The observed performance differences at 54°F outdoor temperature continental interval (HF) between different tests is mainly due to differences in test unit dynamic response in its on/off cycling behavior, indoor temperature variation, and convergence period. In this test interval, UUT showed a mix of regular and irregular on/off cycling behavior which highlights a possible issue related to the test unit's inconsistent dynamic response. Figure 183, Figure 184, and Figure 185 show test unit performance in HF test interval for test P1 at UL, test #10 at UL, and test #6 at PG&E, respectively, in increasing order of measured COP. COP differences in tests P1 and #10 at UL are possibly due to variation in test unit on/off cycling behavior, indoor temperature, and convergence period which seems to be mainly due to the test unit's inconsistent dynamic response with its embedded controller and thermostat at same test conditions. Further in test #6 at PG&E, a different on/off cycling pattern was observed, and the indoor temperature was maintained around an average value of 77.7°F, significantly higher than the thermostat setpoint of 70°F and other tests at UL. This could be due to the thermostat offset setting during test unit setup and is highlighted as a possible testing approach issue. Due to these differences in test unit dynamic response and indoor temperature along with some other possible differences in test unit setup and instrumentation between two labs, measured COP at PG&E was quite high compared to UL tests for this test interval. Overall, poor reproducibility was observed in HF continental interval, with differences in COP from the mean of 5 tests varying from -28.3% to 27.6%.

In 47°F outdoor temperature test interval, where test unit variable speed operation mode was captured during convergence in all tests, differences in COP from the mean of 5 tests varies from -7.9% to 5.7%. This test interval shows the best reproducibility in all continental test intervals for NEEA4. In this test interval also, the indoor temperature in PG&E tests during the convergence period was maintained at around 75.5°F, quite higher than in UL tests where it was maintained at around 70.5°F to 71.5°F, possibly due to the difference in thermostat offset settings. Also, a difference in test unit dynamic response was observed between the two labs, in PG&E tests, UUT provided a higher heating rate to compensate for higher building load corresponding to higher indoor temperature. It should be noted in test P1 at UL, that the test unit was cycling on/off in this test interval, however, convergence was found in variable speed mode before the test unit cycled off. Including a complete on/off cycle in average performance will further reduce the COP in test P1, thus increasing the variation in performance in different tests for HE continental test interval.

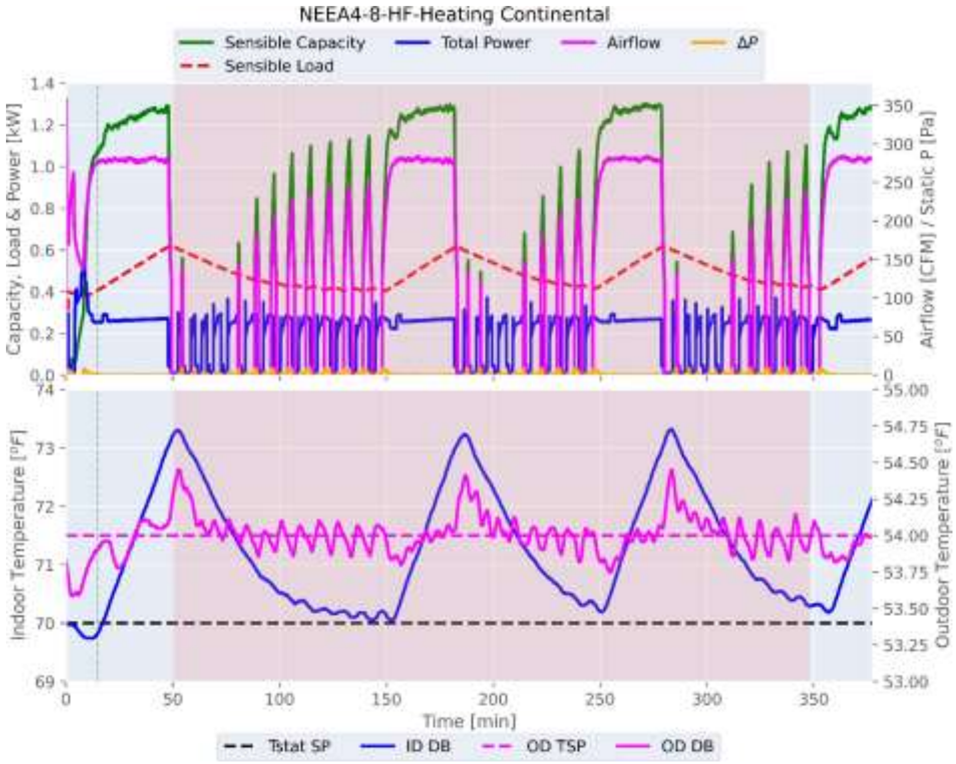


Figure 183. NEEA4-8 (UL) performance and temperature variations for heating continental test interval - HF (54°F)

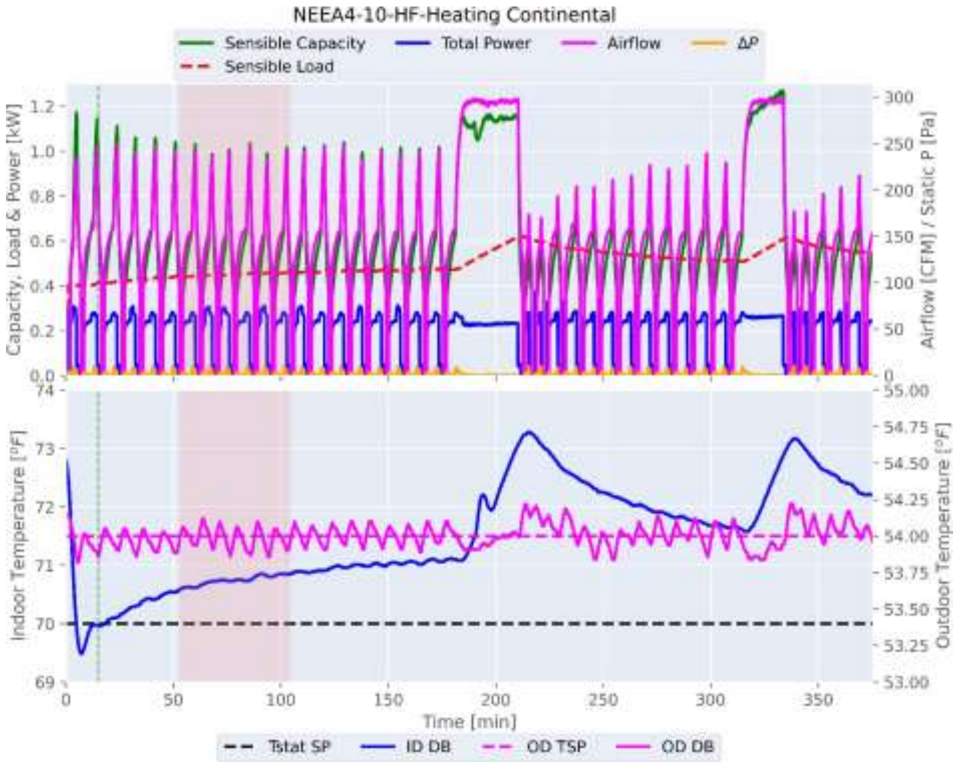


Figure 184. NEEA4-10 (UL) performance and temperature variations for heating continental test interval - HF (54°F)

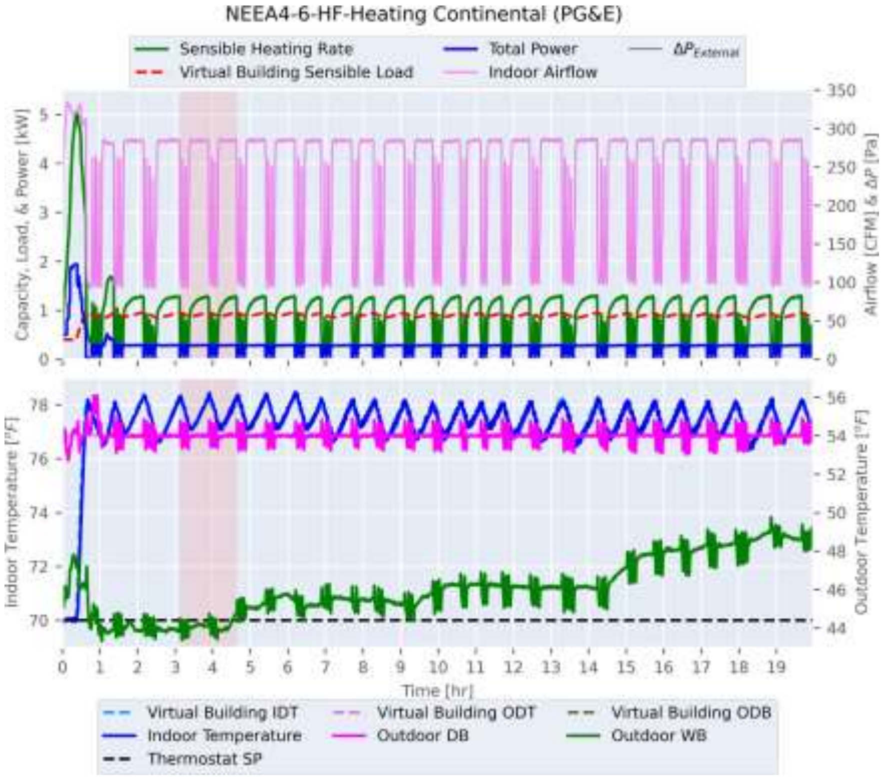


Figure 185. NEEA4-6 (PG&E) performance and temperature variations for heating continental test interval - HF (54°F)

In 34°F ambient temperature continental test interval (HD), poor reproducibility was observed with relatively lower COPs measured in PG&E tests compared to UL tests. Differences in COP from the mean of 5 tests vary from -16.8% to 14.9%. This is possibly due to the difference in test unit dynamic response between the two labs as can be seen in the test unit operation modes captured during convergence. Figure 186 shows the test unit performance in test #10 at UL where the unit was going into defrost, however, convergence was found in variable speed hybrid operation before the unit went into defrost. This shows a possible issue with convergence criteria as for representative performance converged period should include defrost. Also, the test was not run long enough to get multiple complete defrost cycles, highlighting a possible testing approach issue here. Figure 187 shows test unit performance in test #8 at PG&E under the same test conditions where the unit was showing a mix of on/off cycling and defrost and convergence was not found, and 6 hours of data were taken in average performance. In addition to the difference in test unit dynamic response, variation in indoor temperature was also observed in two labs possibly due to thermostat sensing dynamics.

In 17°F and 5°F outdoor temperature test intervals, in general, relatively lower COPs were measured in PG&E tests compared to UL tests with differences in testing approach and test unit operation mode. At UL, load-based tests were performed and convergence was found in variable speed operation mode without including defrosts which shows possible convergence criteria issues for these test intervals. Whereas at PG&E, full-load tests were conducted, in which unit was going into defrost and complete defrost cycles were considered in converged period for test interval average performance. In load-based tests at UL, the indoor temperature varied based on virtual building model response, whereas, at PG&E indoor temperature was kept constant in full-load tests. Due to these differences in test unit dynamic response, testing approach, and indoor temperature, a large variation in test unit performance was observed between the two labs.

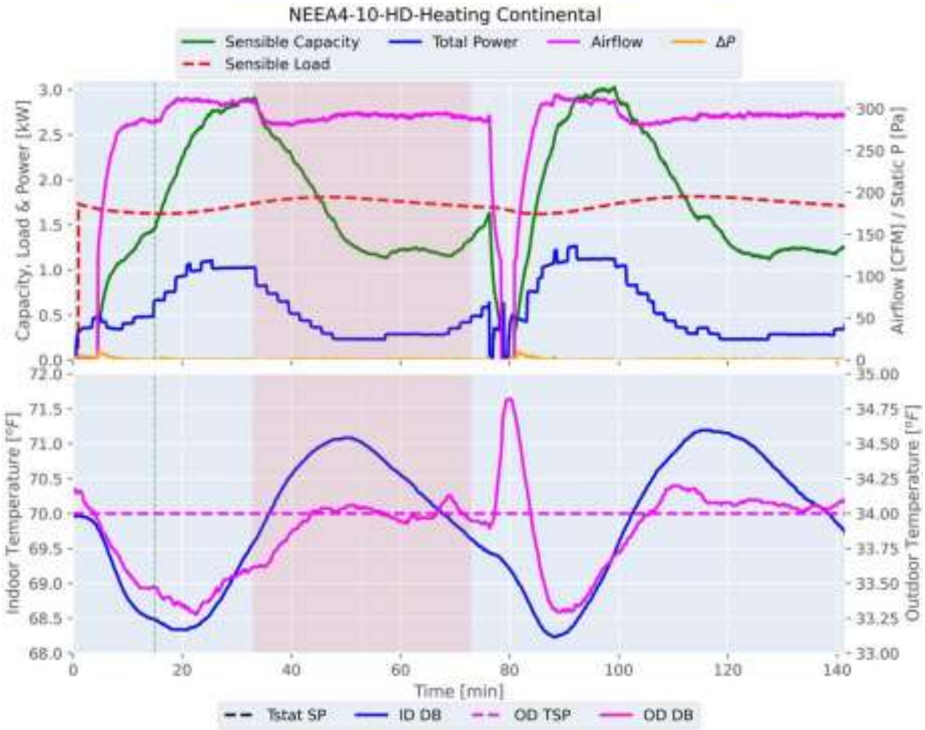


Figure 186. NEEA4-10 (UL) performance and temperature variations for heating continental test interval - HD (34°F)

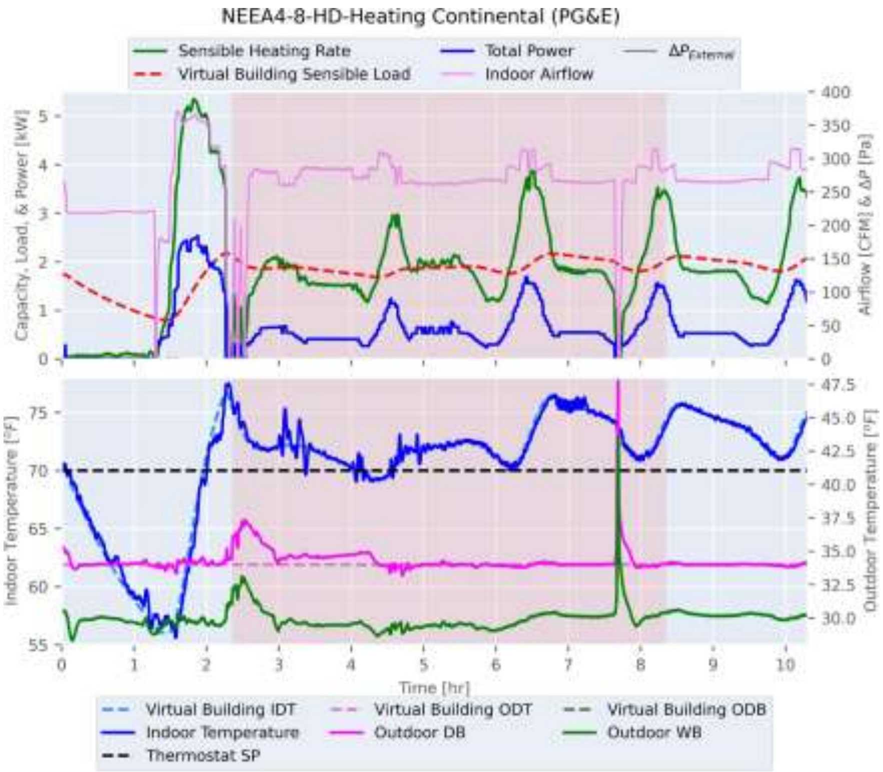


Figure 187. NEEA4-8 (PG&E) performance and temperature variations for heating continental test interval - HD (34℉)

Figure 188 and Figure 189 show the comparison of test unit performance and operation mode during convergence in heating marine test intervals. Possible test interval issues were observed here, which were similar to the continental test interval issues addressed above and in sections 4.2.2 and 4.3.2. Overall similar reproducibility results were observed as continental test intervals at the same ambient temperature test conditions. In 54°F outdoor temperature marine test interval, good repeatability was noted in UL as well as in PG&E tests; however, relatively higher COPs were measured in tests at PG&E compared to UL tests, showing poor reproducibility. This was primarily due to differences in test unit dynamic response with its irregular on/off cycles, relatively higher indoor temperature (6°F-7°F) in PG&E tests, and time where convergence was found.

In marine test interval at 47°F outdoor temperature, where test unit variable speed operation mode was captured during convergence in all 5 tests, relatively better reproducibility was observed compared to other test intervals with the difference in COP from the mean of 5 tests varying from -7.5% to 6.4%. However, still in this test interval, 5°F to 6°F higher indoor temperature was observed in PG&E tests compared to UL tests. But possibly the effect of the difference in indoor temperature might have been counteracted by the difference in test unit dynamic response resulting in comparable COPs in two test labs for this test interval. Further, in the 34°F ambient temperature marine test interval, poor reproducibility was observed with differences in COP from the mean of 5 tests varying from -18.8% to 24.2% where relatively higher COPs were measured at UL. This difference was mainly due to the differences in test unit dynamic response, operation mode captured during convergence, and indoor temperature variation. At UL, convergence was found in variable speed operation mode and did not include defrost in average performance which was only observed in test #10 at UL. Whereas at PG&E, test unit performance did not converge and a mix of on/off and defrost cycles was taken in average performance. Also, the average indoor temperature in PG&E tests was around 1°F to 2°F higher than in UL tests. Overall, this resulted in a quite different test unit performance in two labs at 34°F outdoor temperature.

Furthermore, poor reproducibility in two labs was also observed in the 17°F ambient temperature marine test interval, where around 35% lower COPs were measured in PG&E tests compared to UL tests. However, in each lab, good repeatability was noted. This difference in performance between the two labs was primarily due to the testing approach (load-based vs full load) utilized, test unit dynamic response, and unit operation mode (with or without defrosts) captured in the converged period, similar to continental test interval at the same test conditions.

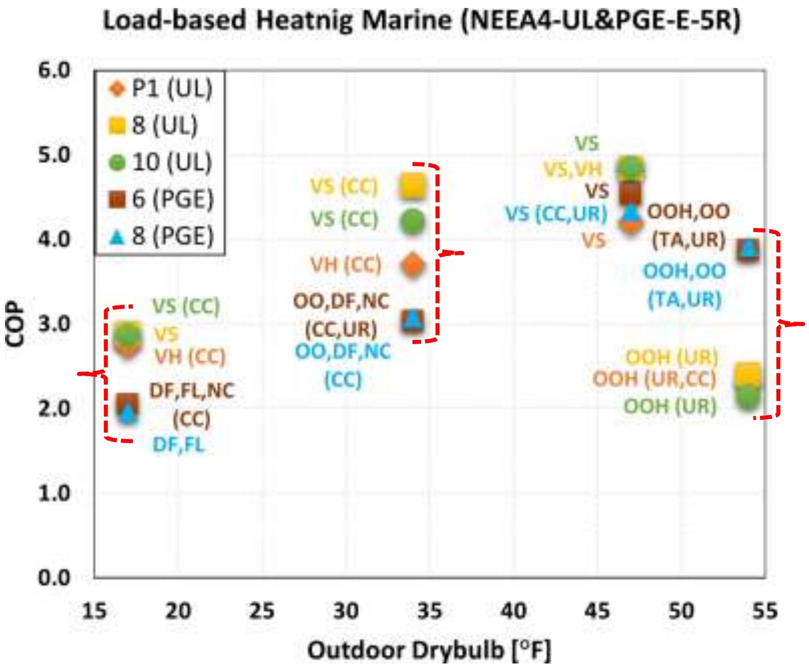


Figure 188. NEEA4 heating marine test intervals COP with unit behavior during convergence

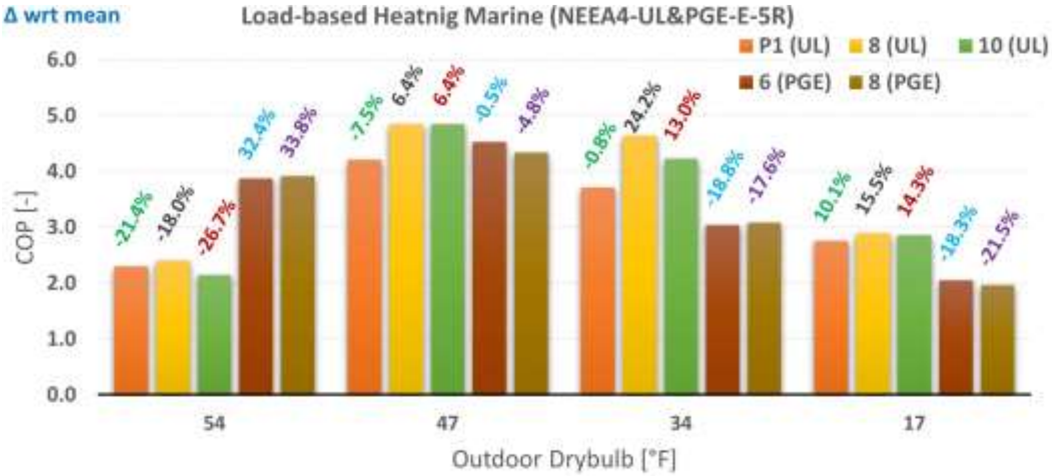


Figure 189. NEEA4 heating marine test intervals COP comparison

Figure 190 shows the overall heating continental and marine test interval results with COP average and standard deviation of 5 tests. In addition, the difference in the minimum and maximum COP for each test interval is shown on the right vertical axis as a percentage of the mean value. It can be seen that overall poor reproducibility was observed in NEEA4 heating tests. For continental test intervals, the maximum difference in COP varies from 13.6% to 55.8% and STD from ±5.4% to ±23.6%. Similar results were observed for marine test intervals in which the maximum difference in COP varies from 13.8% to 60.5% and STD from ±5.7% to ±27.2%.

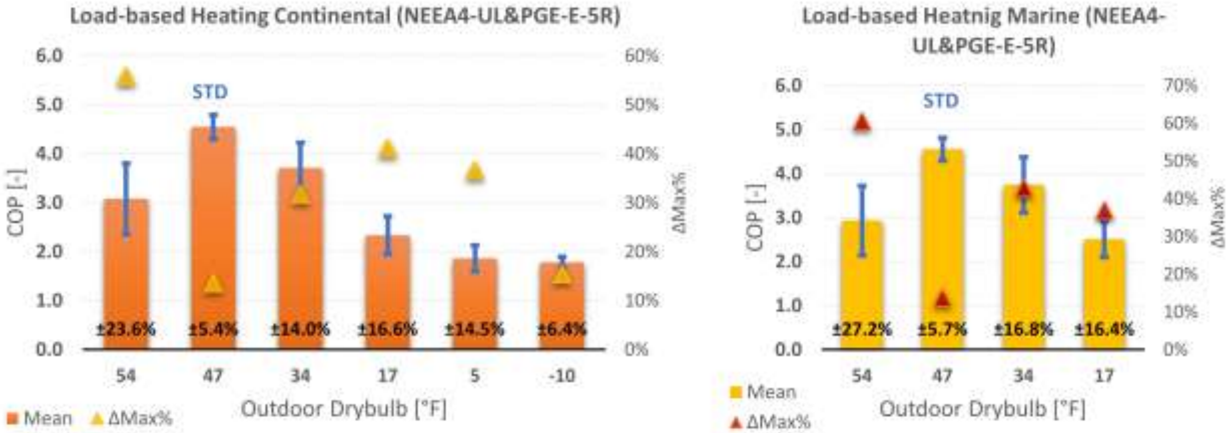


Figure 190. NEEA4 heating continental and marine test interval COP average and maximum difference for different tests with a standard deviation

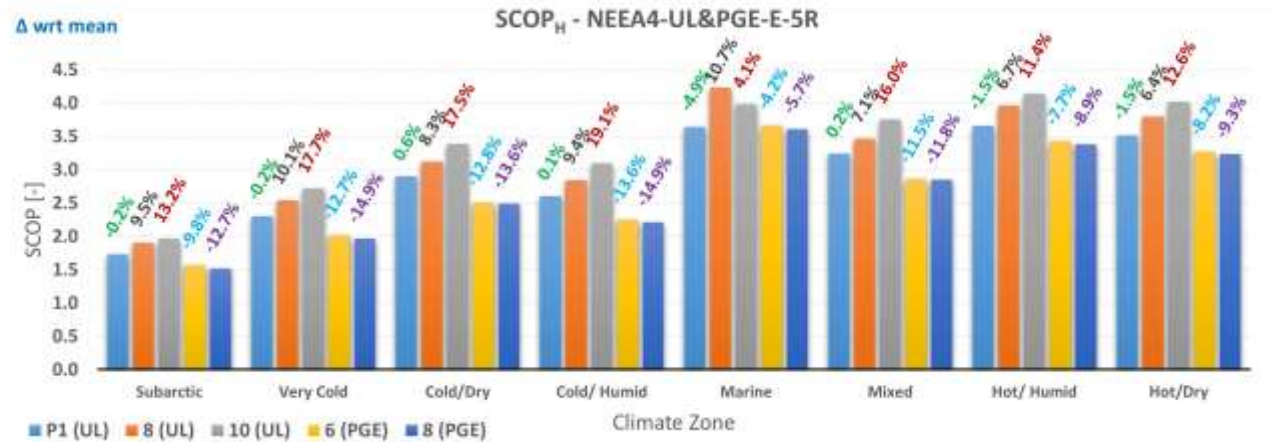


Figure 191. NEEA4 heating SCOP comparison in different climate zones for different tests

Figure 191 and Figure 192 show the comparison of estimated heating SCOP for different climate zones based on 5 sets of test results in two labs along with the 95% confidence interval, standard deviation, and percentage difference between the minimum and maximum SCOP for each climate zone. Overall poor reproducibility was noted across all climate zones, with relatively higher SCOP estimated based on UL test results compared to PG&E. For the "Marine" climate zone which utilizes marine test interval results in SCOP estimation, relatively better reproducibility is noted compared to other climate zones, with a maximum difference in SCOP of 5 tests around 16.5%, 95% CI around ±8.9%, and STD around ±6.4%. For the "Marine" climate zone, maximum SCOP is estimated based on test #8 at UL and minimum based on test #8 at PG&E. For other climate zones for which SCOP is estimated based on continental interval results, the maximum difference in SCOP varies from 20.3% to 34%, 95% CI from ±11% to ±18.2%, and STD from ±7.9% to ±13.1%. For these climate zones, maximum SCOP was estimated based on test #10 results at UL and minimum based on test #8 at PG&E.

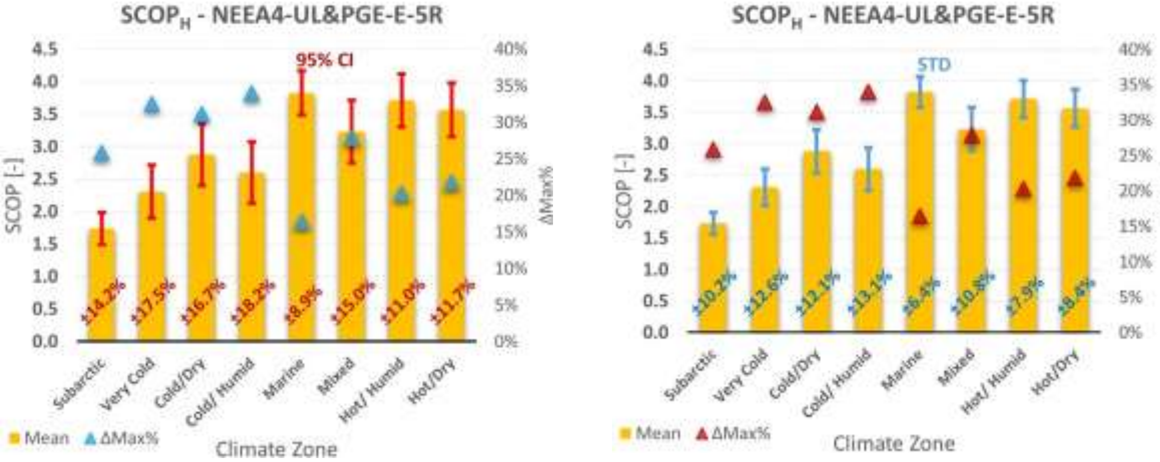


Figure 192. NEEA4 heating SCOP average and maximum difference for repeated tests with a) 95% confidence interval, and b) standard deviation

Table 22 shows the overall NEEA4 SCOP reproducibility results summary with cooling SCOP results crossed out because of the limitation of cooling test results data at PG&E as discussed previously which restricts drawing any meaningful conclusions from cooling SCOP comparison. If comparing only heating results, then overall poor reproducibility was observed with NEEA4 compared to the NEEA7(B) unit as shown in Table 21.

Table 22. NEEA4-UL&PGE-E-5R dataset SCOP reproducibility summary results

Metric	No. of Tests	ΔSCOP (Max-Min) [%]			SCOP 95% CI [%]			SCOP STD [%]		
		Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
~~$SCOP_C$~~	~~5~~	~~31.8%~~	~~39.3%~~	~~35.7%~~	~~19.3%~~	~~21.3%~~	~~20.5%~~	~~13.9%~~	~~15.3%~~	~~14.8%~~
$SCOP_H$	5	16.5%	34.0%	26.3%	8.9%	18.2%	14.2%	6.4%	13.1%	10.2%

6 EXP07 Repeatability and Reproducibility Summary

In this section an overall summary of the EXP07:2019 load-based testing methodology repeatability and reproducibility assessment is provided based on the test results of multiple units in two labs, UL and PG&E, using datasets defined in Table 3. Table 23 presents the repeatability assessment for different datasets showing the minimum, maximum and mean of two statistical parameters across different climate zones for estimated cooling and heating SCOP based on repeated tests. ΔSCOP (Max-Min) is the difference in the minimum and maximum estimated SCOP of different tests for each climate zone described as the percentage of tests' mean SCOP. It represents the maximum variation in performance between different tests. SCOP STD is the standard deviation of SCOP based on repeated tests shown as a percentage of the average SCOP of multiple tests. Figure 193 shows the same results in graphical representation with the mean and range of statistical parameters across different climate zones for different datasets.

Table 23. EXP07 cooling and heating SCOP repeatability assessment summary for UL datasets

			$SCOP_C$			
Dataset	**ΔSCOP (Max-Min) [%]**			**SCOP STD [%]**		
	Min	***Max***	***Mean***	***Min***	***Max***	***Mean***
AFS-UL-E-3R	0.2%	3.0%	1.4%	0.1%	1.2%	0.6%
NEEA4-UL-E-3R	1.5%	14.3%	9.3%	0.7%	6.7%	4.3%
NEEA4-UL-E-2W	0.2%	1.6%	0.8%	0.1%	0.8%	0.4%
NEEA7(B)-UL-E-3R	5.1%	6.8%	5.9%	2.1%	2.8%	2.4%
NEEA7-UL-E-2W	1.5%	3.3%	2.5%	0.7%	1.6%	1.2%
NRCan8-UL-E-2W	0.6%	1.7%	1.1%	0.3%	0.8%	0.6%
NRCan10-UL-E-2W	1.7%	2.7%	2.2%	0.9%	1.3%	1.1%
			$SCOP_H$			
AFS-UL-E-3R	0.6%	1.2%	1.0%	0.3%	0.5%	0.5%
NEEA4-UL-E-3R	12.2%	17.3%	14.6%	5.1%	7.1%	6.0%
NEEA4-UL-E-2W	3.3%	8.5%	6.3%	1.7%	4.3%	3.2%
NEEA7(B)-UL-E-3R	0.8%	4.7%	3.1%	0.3%	2.1%	1.4%
NEEA7-UL-E-2W	0.5%	3.9%	2.7%	0.2%	2.0%	1.3%
NRCan8-UL-E-2W	0.0%	0.8%	0.5%	0.0%	0.4%	0.3%
NRCan10-UL-E-2W	0.1%	1.4%	0.8%	0.0%	0.7%	0.4%

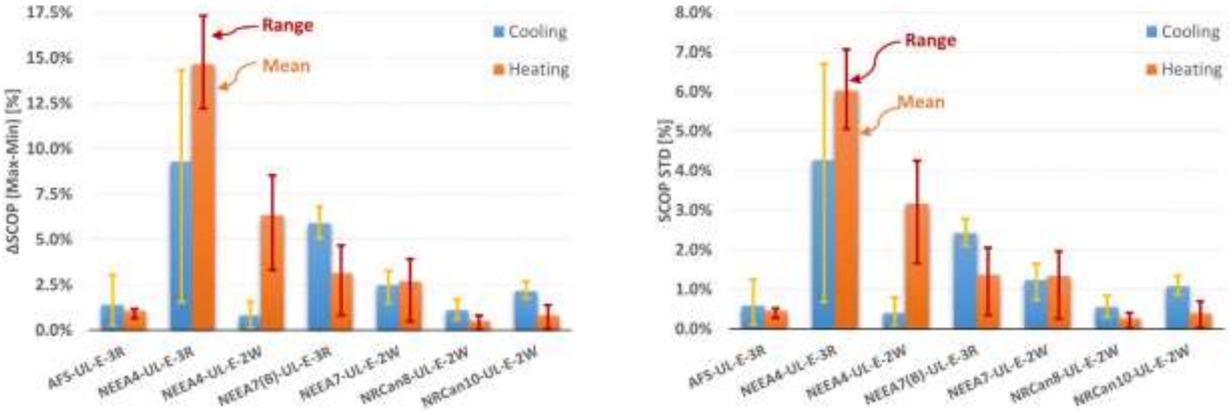


Figure 193. EXP07 cooling and heating SCOP repeatability assessment summary for different datasets from UL with mean and range of statistical parameters across different climate zones

Overall, good repeatability was observed in cooling and heating estimated SCOP of AFS, NRCan8, and NRCan10 test units at UL as shown with results based on datasets AFS-UL-E-3R, NRCan8-UL-E-2W and NRCan10-UL-E-2W. Compared to these three datasets, a relatively larger variation was observed in measured performance among three tests in dataset NEEA7(B)-UL-E-3R with the NEEA7 and NEEA7B unit, where reasonable repeatability was observed with somewhat better results in heating mode compared to cooling mode tests. In heating mode, the maximum difference in estimated SCOP between repeated tests varied from 0.8% to 4.7% and STD from ±0.3% to ±2.1% across different climate zones, whereas, in cooling mode, the maximum SCOP difference varied from 5.1% to 6.8% and STD from ±2.1% to ±2.8%. Furthermore, good repeatability is achieved in both heating and cooling mode when considering the NEEA7-UL-E-2W dataset. That dataset only considers back-to-back tests (P2 and P3) at UL with NEEA7 without including the last test #11 with NEEA7B at UL that was performed after reinstallation about two years later. NEEA4 test results for dataset NEEA4-UL-E-3R showed somewhat poor repeatability, which was mostly due to significantly different measured performance in test P1 that was performed more than a year earlier compared to the other two sets of tests, #8 and #10. Without taking test P1 into account in the NEEA4 repeatability evaluation in dataset NEEA4-UL-E-2W, the test unit shows good repeatability in cooling mode and reasonable repeatability in heating mode. The difference in unit performance during test P1 compared to the other two tests is mostly attributable to differences in test unit dynamic response under the same test conditions and unit dynamic behavior captured during the convergence period in some test intervals as can be seen in section 4.2 of the report. Figure 21, Figure 23 to Figure 25, Figure 26, and Figure 28 to Figure 31 show differences in test unit dynamic response and behavior captured during the convergence period for cooling tests with NEEA4 at UL. Similar differences for heating tests with NEEA4 at UL are shown in Figure 35 and Figure 37 to Figure 45. It is important to note that test P1 with NEEA4 at UL was performed around 16 months before the other two tests, #8 and #10, and the test unit was taken out of the test rooms after test P1 and re-installed. So, in addition to the variability in test unit integrated controls response, differences in NEEA4 dynamic response at the same test conditions could be due to differences in test setup, charge, testing approach, etc. It should also be noted that tests P2 and P3 with NEEA7 were also performed around the same time as test P1 with NEEA4 and around 24 months prior to the 3rd test (#11) with NEEA7B. However, unit NEEA7(B) (NEEA7 & NEEA7B) based on dataset NEEA7(B)-UL-E-3R showed better repeatability than NEEA4 with dataset NEEA4-UL-E-3R based on all three sets of tests at UL.

Table 24. EXP07 cooling and heating SCOP repeatability assessment summary for PG&E datasets

SCOP$_C$						
Dataset	**ΔSCOP (Max-Min) [%]**			**SCOP STD [%]**		
	Min	***Max***	***Mean***	***Min***	***Max***	***Mean***
NEEA4-PGE-E-2W	0.0%	1.2%	0.7%	0.0%	0.6%	0.4%
NEEA7B-PGE-E-2W	0.7%	3.8%	1.9%	0.4%	1.9%	0.9%
SCOP$_H$						
NEEA4-PGE-E-2W	0.4%	3.2%	1.6%	0.2%	1.6%	0.8%
NEEA7B-PGE-E-2W	0.2%	2.1%	0.6%	0.1%	1.0%	0.3%

Table 24 and Figure 194 show the EXP07 repeatability assessment summary results based on datasets of two test units at PG&E. Good overall repeatability was observed for both test units, NEEA4 and NEEA7B, in cooling as well as heating mode. For NEEA4, relatively better repeatability was observed in cooling mode compared to heating mode and the opposite was observed for NEEA7B. In comparison to UL results, PG&E tests results for both units showed better repeatability. This could be because, unlike UL, at PG&E tests were performed back-to-back with no unit re-installation in between. As a result, there was no variability associated with the test setup, refrigerant charge, instrumentation, and other similar factors.

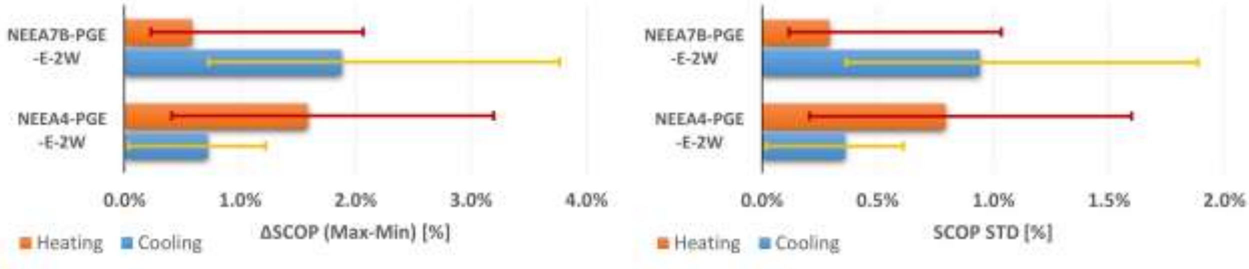


Figure 194. EXP07 cooling and heating SCOP repeatability assessment summary for different datasets from PG&E with mean and range of statistical parameters across different climate zones

Table 25. EXP07 cooling and heating SCOP reproducibility assessment summary for two datasets

SCOP$_C$									
Dataset	ΔSCOP (Max-Min) [%]			SCOP STD [%]			SCOP 95% CI [%]		
	Min	*Max*	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*	*Mean*
~~**NEEA4-UL&PGE-E-5R**~~	~~31.8%~~	~~39.3%~~	~~35.7%~~	~~13.9%~~	~~15.3%~~	~~14.8%~~	~~19.3%~~	~~21.3%~~	~~20.5%~~
NEEA7(B)-UL&PGE-E-5R	8.0%	24.1%	14.6%	3.0%	9.8%	5.8%	4.2%	13.7%	8.1%
SCOP$_H$									
NEEA4-UL&PGE-E-5R	16.5%	34.0%	26.3%	6.4%	13.1%	10.2%	8.9%	18.2%	14.2%
NEEA7(B)-UL&PGE-E-5R	3.0%	4.8%	3.9%	1.2%	1.8%	1.4%	1.7%	2.5%	2.0%

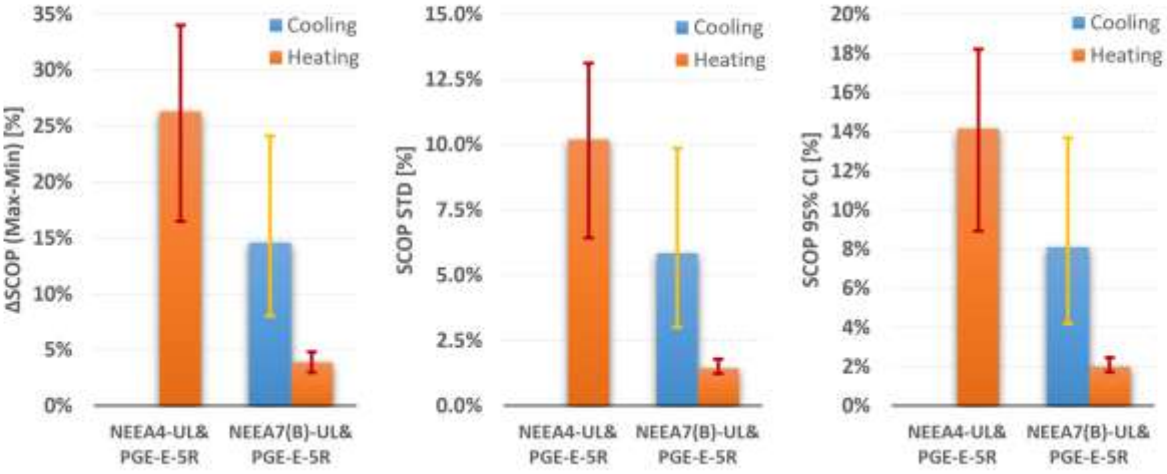


Figure 195. EXP07 cooling and heating SCOP reproducibility assessment summary for two test units for testing at UL and PG&E with mean and range of statistical parameters across different climate zones

Table 25 and Figure 195 show EXP07 reproducibility assessment summary results for two test units with the range and average of statistical parameters across different climate zones based on results from multiple tests at UL and PG&E. In NEEA4 cooling tests at PG&E, full-load tests were performed in some test intervals where load-based tests should have been conducted, due to a possible issue with the testing approach implementation as discussed in section 5.2.1. This resulted in some large differences in cooling SCOP between two labs, and because of this testing approach issue, comparing NEEA4 cooling SCOP between two labs does not provide relevant conclusions regarding EXP07 reproducibility assessment and thus results for this case are crossed out here. Furthermore, heating tests results for NEEA4 showed poor reproducibility with a maximum difference in estimated SCOP between 5 sets of tests for different climate zones varying from 16.5% to 34%, STD from ±6.4% to ±13.1%, and 95% CI from ±8.9% to ±18.2%. Compared to NEEA4, NEEA7(B) demonstrated better reproducibility in both cooling and heating results. Overall good reproducibility was observed in NEEA7(B) heating test results with the maximum difference in estimated SCOP between 5 sets of tests varying from 3% to 4.8%, STD from ±1.2% to ±1.8%, and 95% CI from ±1.7% to ±2.5% across different climate zones. In cooling test results, NEEA7(B) showed poor reproducibility with the maximum difference in estimated SCOP varying from 8% to 24.1%, STD from ±3% to ±9.8%, and 95% CI from ±4.2% to ±13.7% between different tests across various climate zones.

Some of the possible reasons identified for the differences observed in test unit measured performance in different tests in the same lab or different labs at the same test conditions are summarized below.

- **Test unit dynamic response:** In some test intervals, differences in test unit dynamic response along with indoor temperature and humidity variation were observed at the same test conditions which can be due to its inbuilt controller design, hysteresis in control, learning behavior, thermostat sensing response, and thermostat offset. In some cases where a test unit was cycling on/off in different tests at the same test conditions, a difference in test unit cycling rate was observed along with a difference in indoor temperature as well as humidity variation based on virtual building model response (e.g., Figure 137 to Figure 139 and Figure 144 & Figure 145). A difference in the average indoor temperature maintained as well as the range of indoor temperature change during dynamic tests was also observed, possibly due to variation in thermostat sensing dynamics and its offset settings. In some other test intervals with the unit operating in variable speed to compensate for virtual building load, a difference in compressor speed modulation was observed. At same test conditions, compressor operated in variable steady speed mode in some cases whereas in variable hybrid speed mode in other case with compressor speed modulating up and down without compressor cycling off (e.g., Figure 23 to Figure 25). This resulted in a difference in test unit performance as well as indoor temperature response during dynamics tests. In some heating test intervals, especially for NEEA4, a difference in on/off cycling response was observed at the same test conditions with the unit showing a mix of regular and irregular short on/off cycling behavior (e.g., Figure 37 to Figure 39). Also in heating test intervals, for some test conditions, defrost was observed in one test while not in another causing difference in performance (e.g., Figure 42 to Figure 44). Further, for some cases a difference in test unit operation mode was also observed under the same test conditions, especially between two labs, for example, in cooling mode tests, the test unit cycled on/off in one lab while operating at variable speed in another lab (e.g., Figure 140 and Figure 141).

- **Test setup and test facility control:** Other reasons for test unit performance variation at the same test condition between two labs or in the same lab tests with unit re-installation in between is test setup including charge, thermal mass, instrumentation location, sensors dynamic response, etc. For a test unit, a difference in the rate of change of cooling or heating rate can be seen between two labs when the compressor starts up and stops, which is due to the difference in effective thermal mass between test unit supply and supply air side measurements, instrumentation location, and sensor response dynamics (e.g., Figure 138 and Figure 139). The other source of differences between the two labs is the variation in test facility control to maintain test conditions. For example, in some PG&E tests relatively large fluctuations in indoor temperature around virtual building temperature setpoint were observed (e.g., Figure 152 and Figure 154). Also, between the two labs, some differences in outdoor conditions fluctuations when the test unit cycled off or went into defrost were observed which also contributed to variation in overall measured performance (e.g., Figure 162 and Figure 163). Further, a difference in measured airflow was also observed between two labs for the same test unit, with relatively lower airflow measured at PG&E compared to UL at the same test conditions due to a difference in external static pressure control using the booster fan in the airflow measurement device (e.g., Figure 138 and Figure 139). It should also be noted that variation in test unit dynamic response at same test conditions also contributes to airflow differences between two labs in some cases (e.g., Figure 154 and Figure 155). The other factor in the observed airflow differences between the two labs is possibly the difference in fan speed settings during test unit setup. In some tests, possible issues related to external static pressure control were also identified where large fluctuations were observed in measured airflow and external static pressure control (e.g., Figure 132). This requires re-tuning of control parameters for external static pressure control in dynamic load-based tests for some test facilities.

- **Convergence criteria and testing approach:** Another source of differences between test unit measured performance in different tests under the same conditions is the difference in test unit operation mode and dynamic behavior captured in the converged period based on current convergence criteria. These differences are mainly due to issues with the implementation of convergence criteria as per EXP07 as well as some shortcomings in the current convergence criteria. For example, in some test intervals when a test unit was operating in variable hybrid speed mode with the compressor modulating up and down in a steady periodic fashion and convergence was found in a 40-min window, but different dynamic responses were captured in converged periods (e.g., Figure 24 & Figure 25, Figure 76 & Figure 77). In some cases, convergence was found around the valley of power consumption in variable hybrid speed cycles and other cases around the peak, thus resulting in quite different performance even though the overall test unit showed a similar dynamic response. Another example is when a test unit is cycling on/off and/or going into defrost cycles, but in some tests, convergence was found in a 40-minute steady operation period, thus not accounting for on/off cycles and/or defrost cycles in overall average performance for that test interval and resulting in different operation modes as well as average performance captured during convergence at the same test conditions (e.g., Figure 53, Figure 82 to Figure 84, Figure 112). This inconsistency in convergence criteria implementation coupled with variation in test unit dynamic response results in large performance differences between tests for some test intervals, especially between different labs. The other similar convergence criteria implementation issue is when the test unit cycling on/off and/or going in defrost cycles, but convergence is not achieved in the defined maximum test time, but the final converged period takes the average of the entire test duration for overall test interval performance rather than just considering only complete on/off and/or defrost cycles in average performance (e.g., Figure 161 and Figure 162). This also contributes to the test unit measured performance variability at the same test conditions. The other area where current convergence criteria somewhat lack is handling the case when a test unit shows irregular on/off cycling pattern with a mix of short along with regular on/off cycles (e.g., Figure 183 to Figure 185). Due to this, for a test interval, different cycling behavior was captured during the converged period which contributed to performance difference at the same test conditions.

In some test intervals, quite different average indoor temperature was observed during the converged period, especially between two labs, at the same test condition which is possibly due to the difference in thermostat offset settings at the beginning of testing (e.g., Figure 184 and Figure 185). This also contributes to observed performance differences in different tests and is somewhat related to the testing approach to set up correct thermostat offset settings. The other testing approach issue which contributed to performance differences between the two labs is related to the decision to change to a full-load test in a test interval (e.g., Figure 173). A difference in test unit dynamic behavior along with a difference in personnel interpretation of testing approach about when to change to full-load test resulted in full-load tests conducted in one lab whereas load-based tests in another lab at the same test conditions, resulting in measured performance variations. The other possible issue observed with the testing approach is related to the test was not run long enough in some heating test intervals to capture defrost behavior and/or multiple complete defrost cycles as convergence was found in a 40-min window during steady operation mode (e.g., Figure 186). This along with the convergence criteria implementation issue of considering complete defrost cycles resulted in different operation modes captured during convergence in different tests at the same test conditions which resulted in different measured performance.

- **Other:** Additional sources in the measured performance differences at the same test conditions between different tests in the same lab and different labs are uncertainties associated with measurements as well as overall dynamic testing approach, differences in break-in time for each test unit, control, and design airflow settings during test unit setup, etc.

Table 26 provides a categorization of different factors that contribute to the overall repeatability and reproducibility of the EXP07 test results along with a description of their influences and variability.

Table 26. Categorization of variables contributing to EXP07 repeatability and reproducibility

No.	Factor	Repeatability Issues	Reproducibility Issues
1	**Unit under Test (UUT) Installation/Setup** - includes differences in refrigerant charge, duct static pressure setup, thermostat offset setup, type & installation of instrumentation	Relevant when unit is re-installed between tests, especially when installations occur many months apart, when different personnel are involved, and if interpretation of the standard setup changes	An inherent issue for any test standard, but load-based testing has additional installation requirements that are less familiar to personnel and that are still evolving
2	**Environmental Chamber Characteristics** - includes differences in air flow and temperature distribution,	Potentially an issue if a UUT were re-installed within a different chamber between tests or at a different location within the same chamber; also	Differences in air flow and temperature distribution could impact dynamic response of the thermostat; in some cases, chamber

	responsiveness of reconditioning controls	an issue if chamber controls do not track indoor or outdoor setpoints well due to UUT dynamic behavior	controls may not track indoor or outdoor setpoints well due to UUT dynamic behavior
3	**Interpretation and Implementation of Method of Test** - includes convergence criteria, transition to full-load tests	Especially relevant if different personal involved in implementation and/or when repeatability tests are separated over long duration. Also, there are some shortcomings with the current convergence criteria.	An inherent issue for any test standard, but load-based testing has additional requirements that are less familiar to personnel and that are still evolving
4	**Dynamic Behavior of UUT** - covers variations in dynamic behavior of UUT due to variation in control behavior that may result from adaptive learning and limited time period available for testing	Lack of repeatability due to inconsistent control behavior is an inherent issue for some equipment that probably needs to be addressed by each manufacturer	Lack of reproducibility due to inconsistent control behavior is an inherent issue for some equipment that probably needs to be addressed by each manufacturer
5	**UUT Replacement** - due to failure or performance degradation	Performance varies due to manufacturing tolerances and/or firmware differences	Performance varies due to manufacturing tolerances and/or firmware differences

Table 27 provides a qualitative assessment of the cooling and heating mode repeatability results for each unit tested in the two labs along with the possible factors as per Table 26 contributing to the observed differences between repeated tests. The factors considered to be the most important in contributing to observed large differences between different tests are described. As mentioned previously, the UL repeatability results for AFS, NEEA4, and NEEA7(B) based on all three repeated tests at UL, as shown with datasets AFS-UL-E-3R, NEEA4-UL-E-3R, and NEEA7(B)-UL-E-3R are somewhat between true repeatability and reproducibility results because the units were re-installed once in between repeated tests. This could be one of the primary factors for relatively poor repeatability with the NEEA4-UL-E-3R and NEEA7(B)-UL-E-3R datasets at UL compared to other datasets. For both of these units, when only considering two sets of back-to-back tests at UL, in datasets NEEA4-UL-E-2W and NEEA7-UL-E-2W, repeatability improves for both cooling and heating modes as shown in Table 26.

Table 27. EXP07 repeatability assessment summary for different datasets in two labs along with key factors contributing to observed differences

Dataset	Repeatability Assessment		Important Factors
	Cooling	Heating	
AFS-UL-E-3R	Good	Good	**1** - unit was re-installed after being in storage for a period of time and then retested.
NEEA4-UL-E-3R	Poor	Poor	**1, 3, 4** - unit was re-installed after a period of time and then re-tested; convergence criteria implementation; UUT had inconsistent dynamic response and also difference in measured airflow at same test conditions. Refer to Figure 21, Figure 23 to Figure 25, Figure 26, Figure 28 to Figure

			31, Figure 35, Figure 37 to Figure 44, and Figure 45 in section 4.2 for examples.
NEEA4-UL-E-2W	Good	Fair	**3, 4** - differences in dynamic response captured with convergence criteria in different tests; UUT had inconsistent dynamic response at same test conditions. Refer to the same figures as provided in above cell to see some of differences in NEEA4 test #8 and #10 which are included in this dataset.
NEEA7(B)-UL-E-3R	Fair	Good	**1, 3, 4, 5** – unit re-installed after period of time and re-tested; UUT was replaced due to some issues while setting up for testing at PG&E lab; UUT had inconsistent dynamic response; also had differences in dynamic response captured with convergence criteria. Refer to Figure 68, Figure 70 to Figure 73, Figure 74, Figure 76, Figure 77, Figure 80, Figure 82 to Figure 84, and Figure 85 in section 4.4.
NEEA7-UL-E-2W	Good	Good	**3, 4** - some differences in dynamic response captured with convergence criteria in different tests; UUT had inconsistent dynamic response at same test conditions. Refer to the same figures as provided for NEEA7(B)-UL-E-3R to see some of differences in NEEA7 tests P2 and P3 which are included in this dataset.
NRCan8-UL-E-2W	Good	Good	**-**
NRCan10-UL-E-2W	Good	Good	**-**
NEEA4-PGE-E-2W	Good	Good	**-**
NEEA7B-PGE-E-2W	Good	Good	**-**

Table 28. EXP07 reproducibility assessment summary for two datasets along with key factors contributing to observed differences

Dataset	**Reproducibility Assessment**		**Important Factors**
	Cooling	**Heating**	
NEEA4-UL&PGE-E-5R	-	Poor	**1, 2, 3, 4** - differences associated with the facilities and UUT installation/setup including thermostat offset settings and measured airflow; UUT had inconsistent dynamic response; differences in convergence criteria implementation and dynamic response captured during convergence period; also difference in transition to full-load. Refer to Figure 167, Figure 169Figure 170 to Figure 173, Figure 174, Figure 176, Figure 177, Figure 181, Figure 183 to Figure 187, and Figure 188 in section 5.2.
NEEA7(B)-UL&PGE-E-5R	Fair	Good	**1, 2, 3, 4, 5** - differences associated with the facilities and UUT installation/setup; measured airflow; UUT had inconsistent dynamic response; differences in convergence criteria implementation and dynamic response captured during convergence period; also

	🟨	🟩	UUT was replaced in between tests. Refer to Figure 135, Figure 137 to Figure 141, Figure 142, Figure 144, Figure 145, Figure 149, Figure 151 to Figure 158, Figure 159, and Figure 161 to Figure 163 in section 5.1.

It should be clear from the results of Table 27 and Table 28that performance results are much less consistent after equipment has been re-installed in either the same or a different laboratory than if the equipment is retested with the same installation. Some of the important factors that affect the reproducibility include issues related to the installation and setup of the unit for testing (e.g., instrumentation, refrigerant charge, duct static pressure, etc.), different interpretations and implementations of the draft standard, and different characteristics of the environmental chambers. Although these factors are also relevant for the current standard (AHRI 210/240), their impact on the results for load-based testing might be more significant because the draft standard is new and more complicated and the dynamic behavior of the unit with its integrated controls may be sensitive to some installation effects. In addition to the effect of installation and implementation issues on test unit dynamic behavior, variations in test unit dynamic response at the same test conditions could be due to adaptive/learning behavior of the embedded controller and the limited time available for testing. Some of the differences in implementation will likely be resolved as the standard matures and personnel become more familiar with its application. Some of the issues with differences in facilities could be addressed by facility and test equipment improvements. For example, utilizing a thermostat apparatus to provide a standardized environment for the thermostat could reduce differences that are caused by non-uniform airflow and temperatures within environmental chambers. However, differences related to changes in test unit controller behavior would be challenging to address with the test standard without dramatically increasing the testing time. In fact, one of the merits of the load-based testing approach is that it can capture the impacts and sensitivities of test unit performance to dynamic responses of its embedded controller. If a test unit controller performs inconsistently with load-based testing in the laboratory, then it is likely to perform similarly in the field and it is important to capture these effects in a standard method of test. This could be an incentive for manufacturers to develop controllers with more consistent and predictable behavior. With that being said, it would be a good idea to further investigate how best to test and rate units that have inconsistent controller behavior using load-based testing in a repeatable and representative fashion for making the standard more inclusive.

The other possible sources of differences in test unit performance at the same test conditions related to test setup and test facility control can also be addressed by defining appropriate test setup requirements and/or corrections mainly for the components which affect the dynamic performance measurement including code-tester thermal mass, instrumentation location, sensors, etc. Also, appropriate test condition and operating tolerances should be defined to limit the effect of variations in test facility control on overall performance measurement. However, both of these improvements require some further research and investigation as most of the currently established methods of tests are for steady-state tests rather than dynamic load-based tests. Also, the issue with the external static pressure control and airflow measurement difference in the two labs should be further investigated to update the current EXP07 testing approach in order to have better reproducibility in different labs. Moreover, another source of performance variations among different tests is due to some shortcomings in the current convergence criteria and misinterpretation in its implementation as was shown with some example results in the above subsections. Based on the analyzed data results, some points applicable to the current CSA EXP07:19 standard draft convergence criteria are outlined below to improve its repeatability and reproducibility as well as to ensure that representative performance is captured for a test interval.

- Convergence criteria should consider the variable hybrid cycles in which the test unit compressor ramps up and down in a periodic cycling fashion without turning off. In this case, rather than looking for convergence in a 40-minute moving window, the convergence should be checked among these cyclic patterns of peaks and valleys. To make sure that false convergence doesn't occur in the 40-minute moving window, in addition to COP, additional constraints on power and capacity can also be included.
- In a test interval, if the test unit goes into defrost, then convergence criteria should consider the defrost cycle in average performance for that test interval.
- In a test interval, if the test unit cycles on/off, then convergence criteria should consider the on/off cycle in the average performance measurement of that test interval.
- If a test unit shows on/off cycling behavior or defrost cycling behavior and convergence is not achieved in maximum test duration, then in average performance only complete on/off or defrost cycles should be considered. In this case of non-convergence if a test unit is showing both on/off cycles with intermitted defrost cycles, then it should be defined which complete cycles to consider in average performance. Maybe the cycles which span the longer time duration in the test interval can be considered.
- Further investigation should be done on how to consider inconsistent on/off cycling behavior, e.g., a mix of short cycles with longer regular cycles, in convergence criteria to ensure repeatable and reproducible test results
- It might make sense to include a minimum test time in the convergence criteria to observe whether UUT cycles off or goes into defrost during a test interval even if the test is converged based on another criterion such as a 40-minute steady moving period. This could be defined only for test intervals in which the test unit is expected to show on/off cycling or defrost behavior to not unnecessarily increase the testing time.
- Further investigation into including indoor RH variation in convergence criteria and defining bounds in which a test unit should maintain indoor RH during cooling humid coil test intervals for it to be acceptable for thermal comfort conditions.
- Updating and clarifying the approach to decide when to change to a full-load test to address the issues identified previously regarding this between two labs. Also, it might be a good idea to have a provision and approach to double-check whether the decision to change to a full-load test was right or not, maybe comparing the building load and maximum cooling or heating rate during the full-load test.

Some or all the above-outlined recommendations and proposed changes with convergence criteria are possibly addressed in the latest updates to EXP07. In that case, it is recommended to re-analyze the raw test data with updated convergence criteria to assess and ensure whether the updates resolve the possible issues observed here and improve the repeatability and reproducibility or not. Further, some additional areas of consideration are outlined below to improve the load-based testing methodology and its implementation based on the observations in this study.

- In some test intervals, a large difference in average indoor temperature was observed in two labs at the same test conditions, which was possibly due to thermostat offset settings. So, it might be worthwhile to further investigate and update or clarify the testing approach to set thermostat offset during test unit set up to have consistency in measured performance in different labs.
- In two labs, a difference in the unit measured airflow as well as external static pressure was observed which warrants further examination into test unit setup from airflow perspective, test facility external static pressure control approach, and required tolerances to ensure reasonable variability in measured airflow between different labs at same test conditions
- Specifying an approach to provide a standardized environment for the thermostat during load-based testing.
- The appropriate time required for a test unit break-in and learning cycle should be considered for future updates through interaction with different manufacturers.
- Further investigation into the test condition tolerances should be carried out to ensure test result repeatability and reproducibility within the acceptable confidence interval among different labs.
- The optimum value of different time parameters in the convergence criteria should be investigated using the available test data with the goal of improving repeatability and reproducibility.
- Approaches for dealing with test units having inconsistent control behavior should be investigated from a repeatability and reproducibility perspective.
- Investigation into the dynamic measurements uncertainties as well as overall uncertainty in dynamic tests due to defined test conditions and operating tolerances to define appropriate confidence interval expected in seasonal performance metric.

7 AHRI 210/240 Repeatability Assessment

In this section, AHRI 210/240-2023 repeatability assessment results are presented based on the test results of three different units in two labs, UL and PG&E, as per defined datasets in Table 3. All three units, AFS, NEEA4, and NEEA7(B) (NEEA7 & NEEA7B), were tested at UL, whereas, only two of these, NEEA 4 and NEEA7B were tested PG&E. In the subsections below, the repeatability assessment of each unit is presented based on test results from two labs. First, each test interval COPs are compared between different tests, and then seasonal energy efficiency metrics, SEER and HSPF, are compared to assess overall repeatability.

7.1 AFS (UL)

In this section, AHRI 210/240 repeatability assessment results for AFS, a 5-ton fixed-speed heat pump, are presented based on two back-to-back repeated sets of tests at the UL lab in dataset AFS-UL-A-2W. Figure 196 shows the COP comparison for cooling test intervals between two tests #3 and #5. Overall good repeatability was observed with differences in COP varying from 0% to 1.1% and STD from ±0% to ±0.4%. For this unit, the optional test was also performed to measure the cycling degradation coefficient (C_D), which was around 0.067 based on test #3 results and around 0.078 based on test #4, quite smaller than the default value of 0.25 in AHRI 210/240.

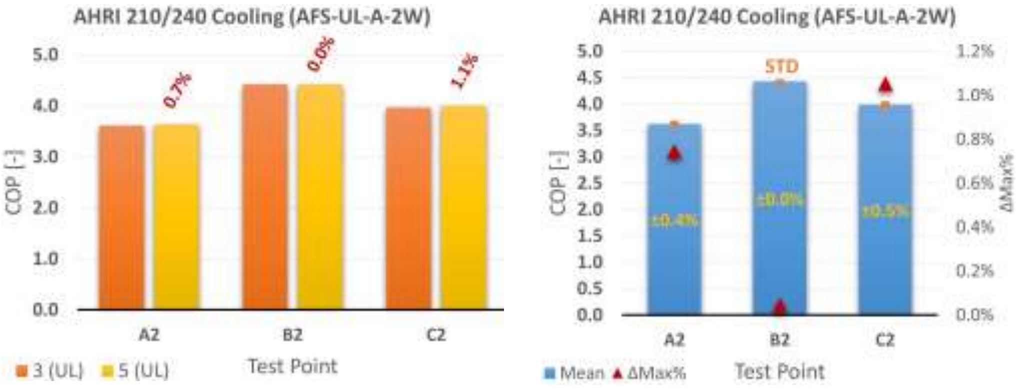


Figure 196. AFS (UL) cooling test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

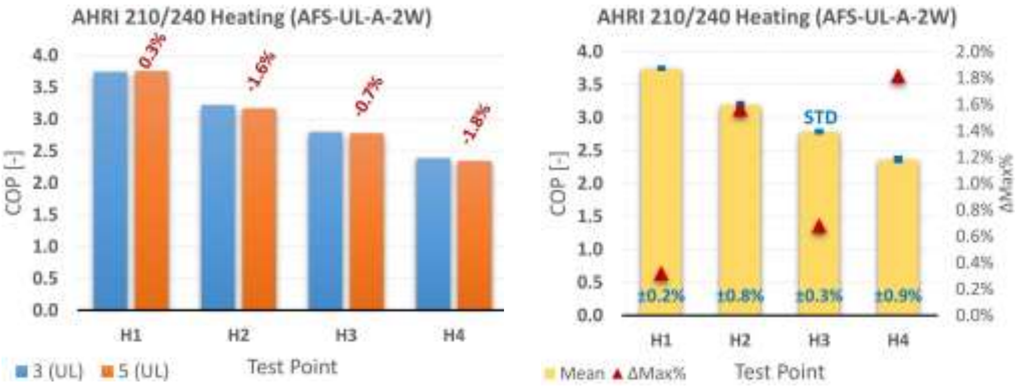


Figure 197. AFS (UL) heating test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Good repeatability was also observed in heating test results as shown in Figure 197, with differences in COP in two tests varying from -1.8% to 0.3% and STD from ±0.2% to ±0.8%. Table 29 shows the comparison of estimated SEER and HSPF with the difference between two tests, and 95% CI based on *t*-distribution and STD as a percentage to mean value. Here overall good repeatability is observed with slightly lower performance based on test #5 results compared to test #4.

Table 29. AFS-UL-A-2W dataset AHRI 210/240 seasonal performance repeatability summary

Metric	AFS-3	AFS-5	∆21%	95% CI	STD
SEER2	14.57	14.55	-0.2%	±1.0%	±0.1%
HSPF2	8.32	8.23	-1.1%	±7.1%	±0.6%

7.2 NEEA4 (UL)

Figure 198 shows the cooling COP comparison for NEEA4, a 1-ton mini-split ductless heat pump, in different test intervals between tests #7 and #9 at UL in dataset NEEA4-UL-A-2W. Higher COPs were measured in test #9 intervals with differences between two tests varying from 0.5% to 12.2% and STD from ±0.3% to ±6.1%. The largest difference was noted in the B1 test interval, in which around a 13.1% larger cooling rate was measured in test #9 compared to test #8 with similar power consumption, airflow, and test room conditions. Further, in heating test intervals, relatively higher COPs were also measured in test #9, as shown in Figure 199, with the difference between the two tests varying from 0.3% to 7.7% and STD from ±0.1% to ±3.8%. The largest differences were noted in test intervals H01 and H11. In the H11 test interval, in test #9, around 10.9% higher heating rate was measured compared to test #7, but with only 2.9% higher power consumption and around 2% (9 CFM) lower airflow rate. As in between these two tests, one set of load-based tests was conducted as per EXP07, so these differences between the two tests could be due to differences in test unit break-in period in addition to test and measurement uncertainties.

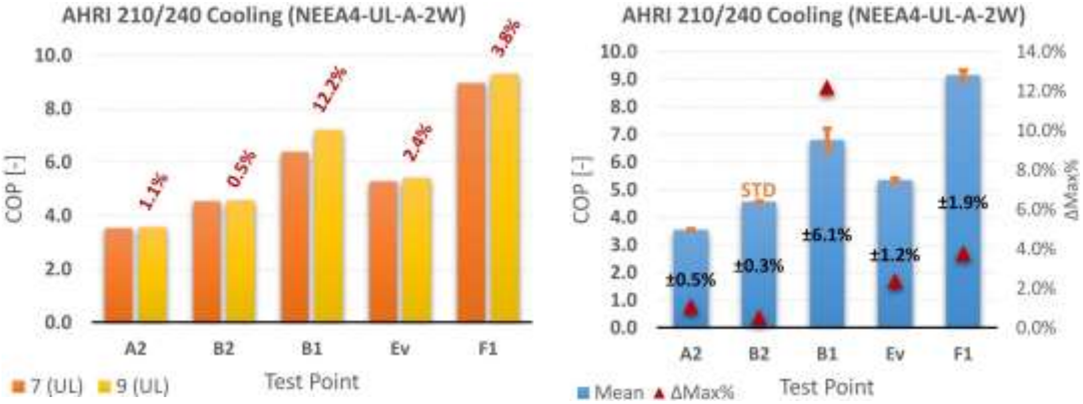


Figure 198. NEEA4 (UL) cooling test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Table 30 shows the comparison of estimated cooling and heating seasonal performances based on these two tests. Relatively poor repeatability was observed in SEER, with a difference of around 7.7% between two tests and STD of around ±3.8%. A relatively larger value of confidence interval was estimated because only two samples were used here. Estimated HSPFs are quite comparable, with a difference of around 1.2% and STD around ±0.6% between the two tests.

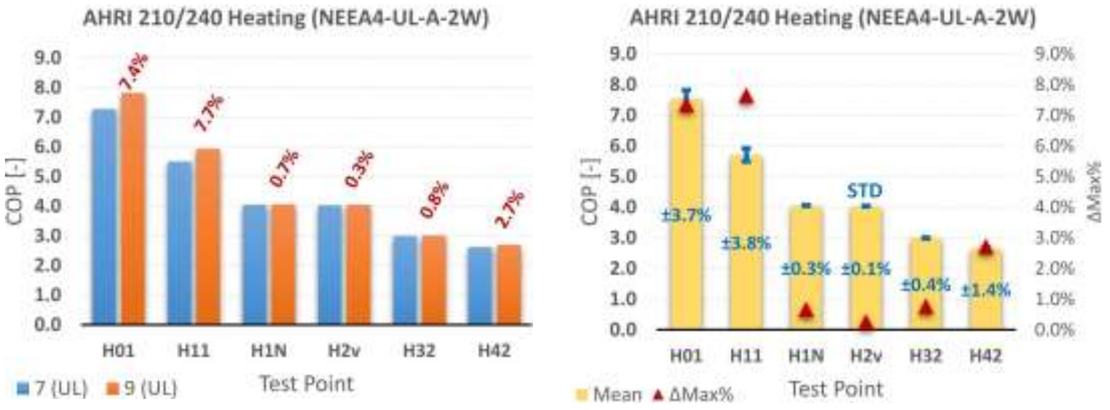


Figure 199. NEEA4 (UL) heating test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Table 30. NEEA4-UL-A-2W dataset AHRI 210/240 seasonal performance repeatability summary

Metric	NEEA4-7	NEEA4-9	∆21%	95% CI	STD
SEER2	19.56	21.12	7.7%	±48.8%	±3.8%
HSPF2	11.13	11.27	1.2%	±7.8%	±0.6%

7.3 NEEA4 (PG&E)

Figure 200 and Figure 201 show NEEA4 COP comparisons between two sets of tests at PG&E, #5 and #7, in dataset NEEA4-PGE-A-2W for cooling and heating test intervals, respectively. Overall test results showed good repeatability, comparatively better than at UL for the same test unit. For cooling, differences in COP between two tests for different test intervals varied from -1.5% to 2.7% and STD from ±0.5% to ±1.4%. Similar repeatability was observed in heating test intervals with differences in COP varying from -2.5% to 0.8% and STD from ±0% to ±1.2%.

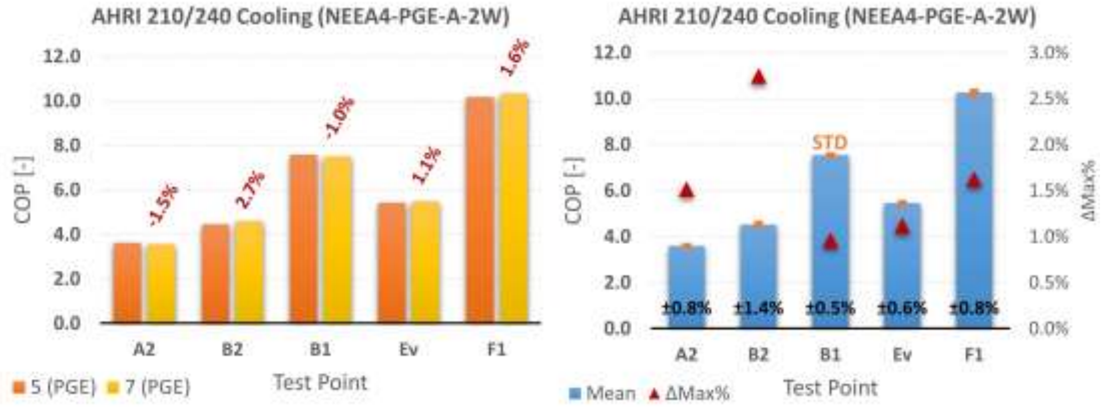


Figure 200. NEEA4 (PG&E) cooling test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

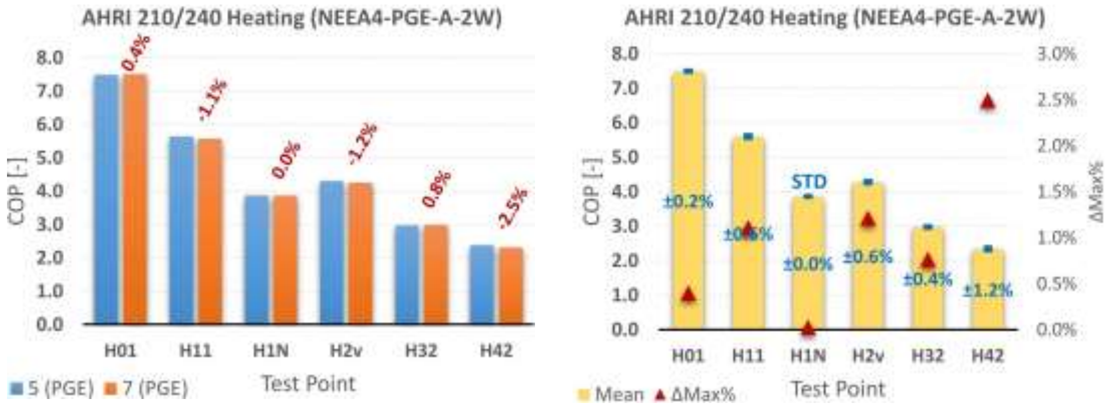


Figure 201. NEEA4 (PG&E) heating test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Table 31 shows the comparison of estimated SEER and HSPF based on measured performance in two tests, showing overall good repeatability.

Table 31. NEEA4-PGE-A-2W dataset AHRI 210/240 seasonal performance repeatability summary

Metric	NEEA4-5	NEEA4-7	∆21%	95% CI	STD
SEER2	21.93	21.88	-0.2%	±1.4%	±0.1%
HSPF2	11.09	11.03	-0.5%	±3.3%	±0.3%

7.4 NEEA7 (UL)

AHRI 210/240 cooling test intervals COP comparison for NEEA7, a 3-ton split-type heat pump, is shown in Figure 202 between 2 tests, #1 and #2, at UL in dataset NEEA7-UL-A-2W. As can be seen, that test unit overall showed good repeatability with differences in COP varying from -2.5% to 2.7% and STD from ±0.2% to ±0.3%.

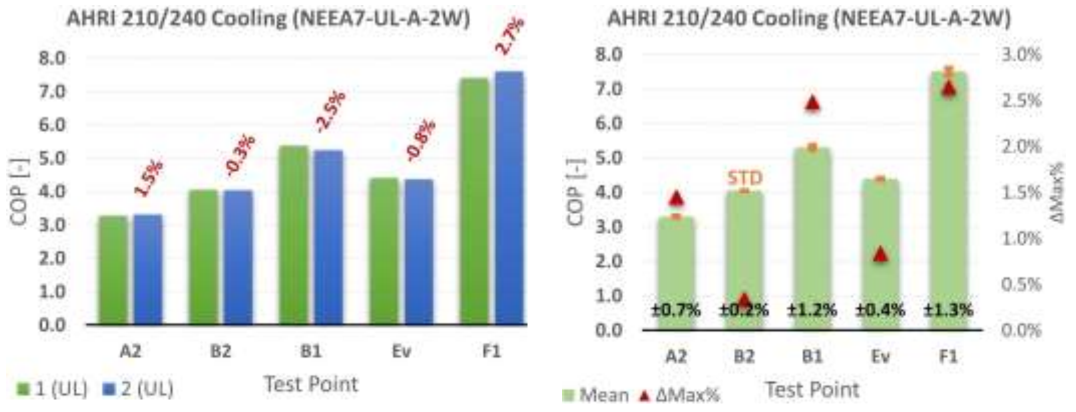


Figure 202. NEEA7 (UL) cooling test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Further, good repeatability was also observed in heating test intervals except for the H32 test interval as shown in Figure 203. The difference in COP between the two tests varies from -1.8% to

4.5% and STD from ±0% to ±2.3%. In the H32 test interval, around 15.2% higher heating rate was measured in test #2 compared to test #1 with around 10.7% higher power consumption, resulting in around 4.5% higher COP. In this test interval, airflow and test room conditions were similar between the two tests.

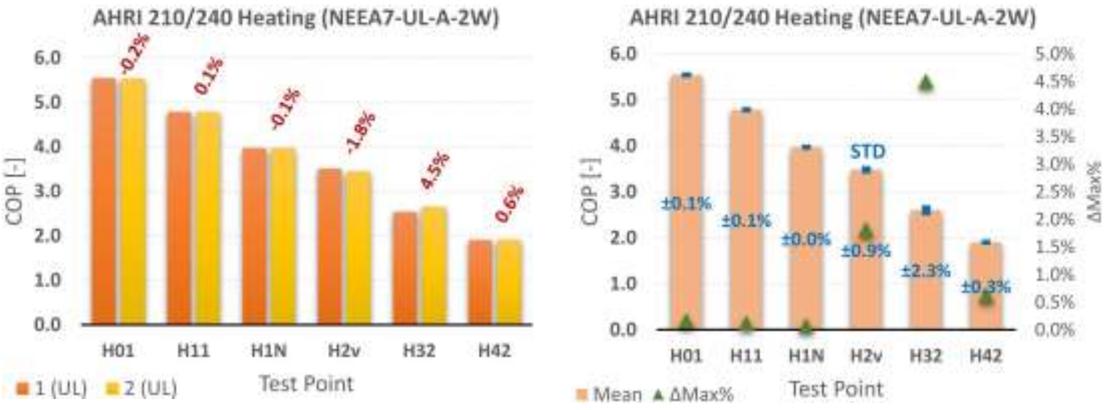


Figure 203. NEEA7 (UL) heating test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Overall good repeatability was observed in seasonal performance metrics, as shown in Table 32 with relatively better repeatability in SEER compared to HSPF. A larger difference in HSPF between the two tests is mainly due to a larger performance difference in the H32 test interval.

Table 32. NEEA7-UL-A-2W dataset AHRI 210/240 seasonal performance repeatability summary

Metric	NEEA7-1	NEEA7-2	Δ21%	95% CI	STD
SEER2	16.03	16.10	0.4%	±2.7%	±0.2%
HSPF2	9.11	9.45	3.7%	±23.5%	±1.8%

7.5 NEEA7B (PG&E)

Figure 204 compares the AHRI 210/240 cooling test intervals COP between two tests, #1 and #3, in dataset NEEA7B-PGE-A-2W from PG&E with the NEEA7B test unit. Across different test intervals except for B1, the difference in COP varied from -1.6% to 1% and STD from ±0.4% to ±0.8%, showing good repeatability. In test interval B1, around 8.5% higher COP was measured in test #3 compared to test #1. In test #3 around 1.4% lower cooling rate was measured, but with around 9.8% lower power consumption, resulting in relatively higher COP compared to test #1 at a similar test condition. Also in test #3 airflow rate was around 50 CFM (6.4%) compared to test #1 which could have been the cause for this performance difference. However, a similar airflow difference between the two tests was also observed in test intervals Ev and F1, but comparable COPs were measured. On the other hand, overall good repeatability was noted in all the heating test intervals as shown in Figure 205. For different test intervals, the differences in COP between the two tests varied from -2.5% to 0.7% and STD from ±0.2% to ±1.3%.

A comparison of estimated seasonal performance metrics based on two tests is shown in Table 33. Around 3.8% higher SEER was estimated based on test #3 results compared to test #1, mainly due to better performance in test #3 B1 interval. Comparatively better repeatability was noted in HSPF, with a difference of around 0.5% between the two tests.

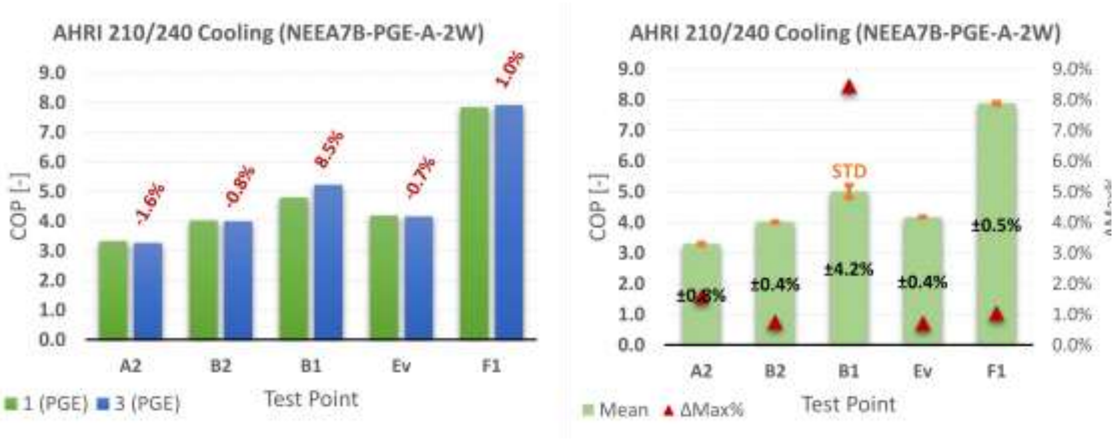


Figure 204. NEEA7B (PG&E) cooling test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

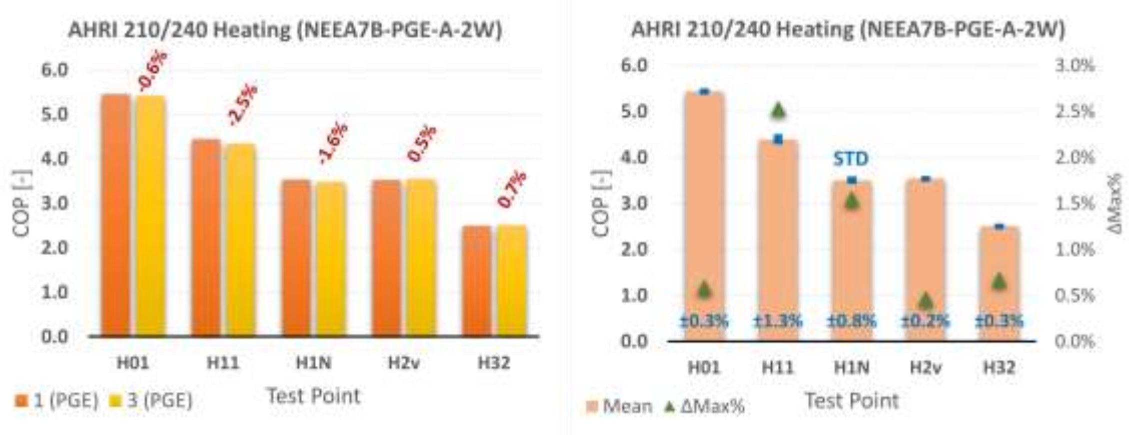


Figure 205. NEEA7B (PG&E) heating test results a) COP comparison in different tests b) Test intervals mean COP with STD and maximum difference among repeated tests

Table 33. NEEA7B-PGE-A-2W dataset AHRI 210/240 seasonal performance repeatability summary

Metric	NEEA7B-1	NEEA7B-3	∆21%	95% CI	STD
SEER2	15.50	16.10	3.8%	±24.2%	±1.9%
HSPF2	9.54	9.50	-0.5%	±3.1%	±0.2%

8 AHRI 210/240 Reproducibility Assessment

In this section, AHRI 210/240 reproducibility assessment is presented based on the test result analysis of two units, NEEA4 and NEEA7(B), for datasets NEEA4-UL&PGE-A-4R and NEEA7(B)-UL&PGE-A-4R. Each unit was tested in two labs, UL and PG&E, twice based on AHRI 210/240-2023. In subsections below, both test units' performance reproducibility analysis is presented, first comparing the measured COP in different tests, and then estimated seasonal performance metrics, SEER and HSPF, are compared.

8.1 NEEA4

For NEEA4, first, two sets of tests (#7 and #9) in dataset NEEA4-UL&PGE-A-4R were performed at UL, then the unit was shipped to PG&E, where the other two sets of tests (#5 and #7) were performed based on AHRI 210/240. Figure 206 shows the comparison of cooling test intervals COP among these 4 tests in two labs. Across these 5 test intervals, the maximum difference in measured COP among different tests varies from 2.5% to 16.7% and STD from ±0.9% to ±6.6%. The largest differences in performance were observed in test intervals B1 and F1 with relatively better performance measured at PG&E. In the B1 test interval, relatively lower cooling rates were measured in UL tests compared to PG&E tests but with similar power consumption, thus resulting in lower COP in UL tests. Further, in this test interval, measured airflow in tests #7 and #9 at UL was around 552 CFM and 547 CFM, whereas in PG&E tests #5 and #7 it was around 469 CFM. This higher airflow in the UL lab could be a possible reason for lower COP in UL tests, however, it seems counterintuitive. Further, a similar difference in airflow between the two labs was also observed in other cooling test intervals as shown in Figure 207, however comparable measured COPs in the two labs. Similarly, in test interval F1, higher COPs in PG&E tests were mainly due to higher cooling rates with similar power consumption compared to UL tests. Also, it's interesting to note that both of these test intervals with poor reproducibility correspond to minimum compressor speed test intervals. So, the observed differences in the two labs could be due to a difference in some control settings during testing.

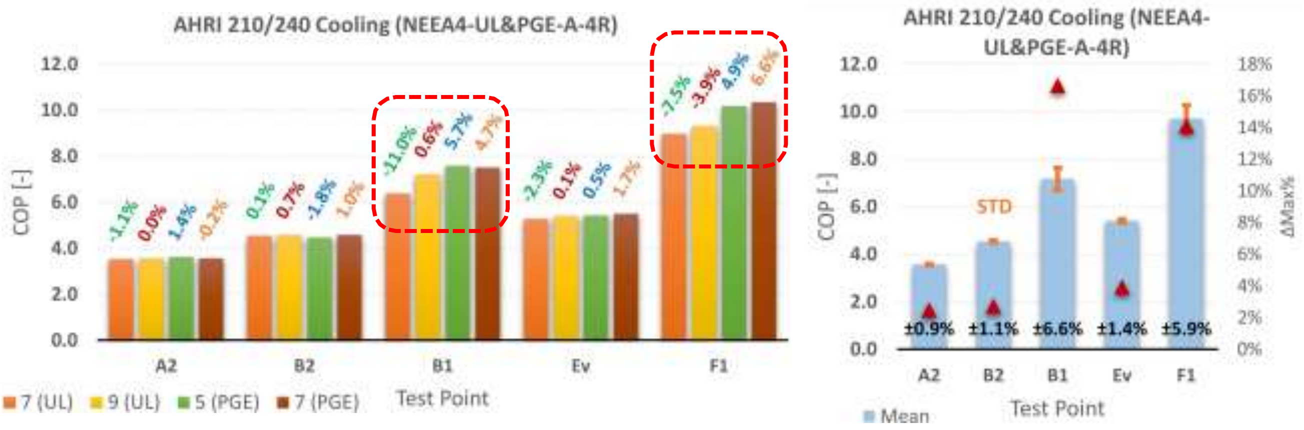


Figure 206. NEEA4 (UL&PGE) cooling test results a) COP comparison b) Test intervals mean COP with STD and maximum difference for AHRI 210/240 reproducibility assessment

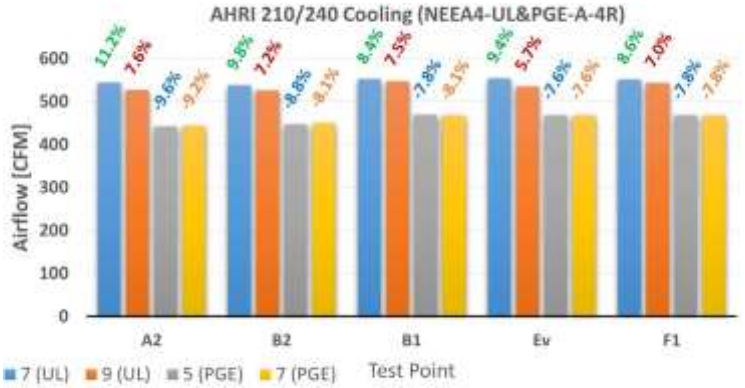


Figure 207. NEEA4 (UL&PGE) cooling test intervals airflow rate comparison in different tests

For heating test intervals, COP comparison between 4 tests is shown in Figure 208, with the maximum difference in COP between different tests varying from 1.3% to 14.8% and STD from ±0.5% to ±6.2%. The largest difference in performance between the two labs was observed in optional test interval H42, with relatively lower COPs measured in PG&E tests compared to UL tests. In this test interval, a relatively lower heating rate was measured at PG&E compared to UL, but with similar power consumption, resulting in lower COP in PG&E tests. Further, measured airflow in UL tests was around 385 CFM, whereas, in tests #5 and #7 at PG&E, it was around 372 CFM and 394 CFM, respectively, showing no clear correlation with differences in COP and measured CFM. In H01 and H11 test intervals, observed variability in measured performance was mainly due to larger differences observed in UL tests rather than differences between the two labs. Also, in heating test intervals, similar to cooling tests, relatively lower airflow was measured in PG&E tests compared to UL except for H32 and H42 test intervals. Overall, the observed differences in the two labs could be due to test unit setup, charge, airflow measurement, external static pressure control, and instrumentation in addition to measurement uncertainties.

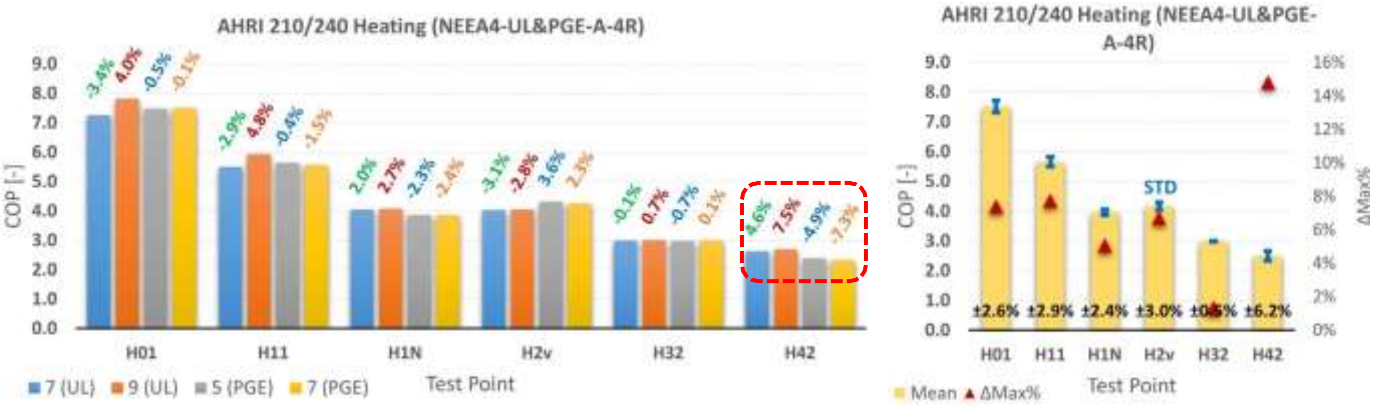


Figure 208. NEEA4 (UL&PGE) heating test results a) COP comparison b) Test intervals mean COP with STD and maximum difference for AHRI 210/240 reproducibility assessment

Table 34 shows the comparison of estimated SEER and HSPF based on these 4 sets of tests along with percentage difference in minimum and maximum value, 95% CI, and STD. For cooling, relatively poor reproducibility was observed with a maximum difference in estimated SEER around 11.2% and 95% CI around ±8.3%. This was mainly due to relatively low estimated SEER based on UL test results compared to PG&E, especially for test #7 at UL. In contrast for heating, the estimated HSPF was slightly higher based on UL test results and overall showed good reproducibility with a maximum difference in HSPF of around 2.2% and 95% CI around ±1.5%.

Table 34. NEEA4-UL&PGE-A-4R dataset AHRI 210/240 seasonal performance reproducibility summary

Metric	7 (UL)	9 (UL)	5 (PGE)	7 (PGE)	∆ [Max - Min] %	95% CI	STD
SEER2	19.56	21.12	21.93	21.88	11.2%	±8.3%	±4.5%
HSPF2	11.13	11.27	11.09	11.03	2.2%	±1.5%	±0.8%

8.2 NEEA7(B)

NEEA7(B) (NEEA7 and NEEA7B) performance was also measured as per AHRI 210/240 in four sets of tests at UL and PG&E in dataset NEEA7(B)-UL&PGE-A-4R. It should be noted that first NEEA7 performance was measured at UL in tests #1 and #2, then with a same model of the unit, NEEA7B, two other sets of tests were conducted at PG&E, #1 and #3. Figure 209 shows the COP comparison of NEEA7(B) cooling test intervals, with the maximum difference in COP between four tests varying from 1.4% to 11.1% and STD from ±0.5% to ±4.2%. The largest difference was noted in the B1 test interval, with minimum COP measured in test #1 at PG&E and maximum in test #1 at UL. In test intervals Ev and F1, the maximum difference in COP between different tests was around 5.9% and 6.7%, respectively. In test interval Ev, relatively lower COPs were measured in PG&E tests compared to UL, whereas in test interval F1 the opposite is observed. Further, it is interesting to note that, in general, a relatively lower cooling rate, power consumption, and airflow rate was measured in cooling test intervals at PG&E compared to UL as shown in Figure 210, Figure 211, and Figure 212. This difference in performance between the two labs could be possibly due to differences in the test setup, instrumentation, and testing approach e.g., control settings. Lower airflows in PG&E tests could be mainly due to higher external static pressure compared to UL tests.

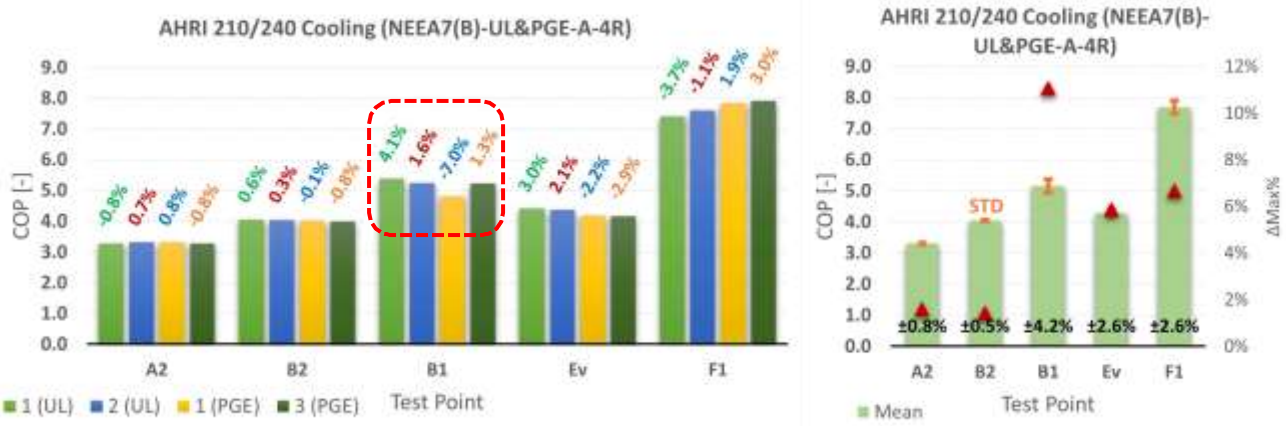


Figure 209. NEEA7(B) (UL&PGE) cooling test results a) COP comparison b) Test intervals mean COP with STD and maximum difference for AHRI 210/240 reproducibility assessment

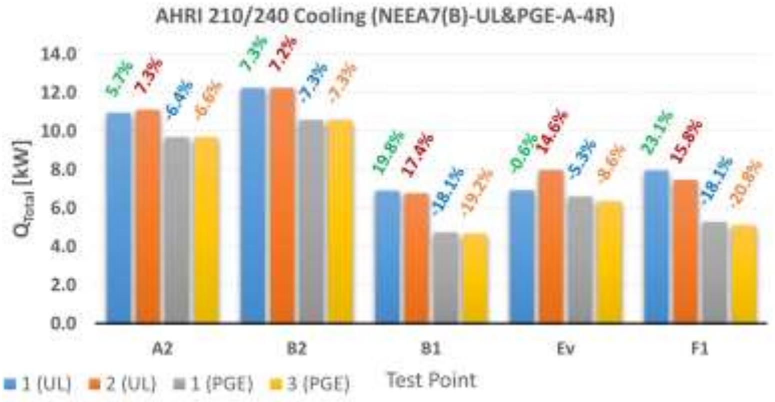


Figure 210. NEEA7(B) (UL&PGE) cooling test intervals cooling rate comparison in different tests

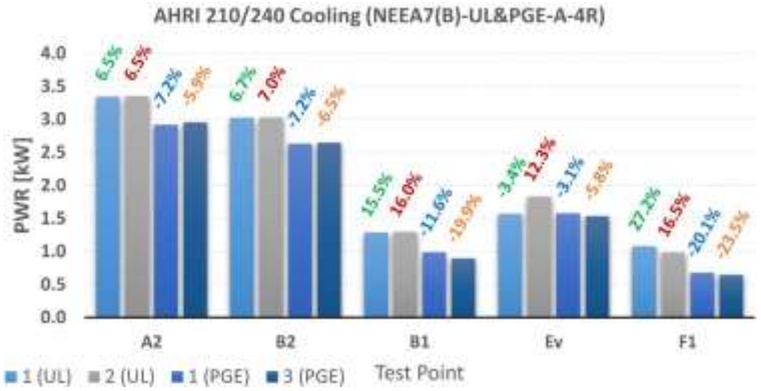


Figure 211. NEEA7(B) (UL&PGE) cooling test intervals power consumption rate comparison in different tests

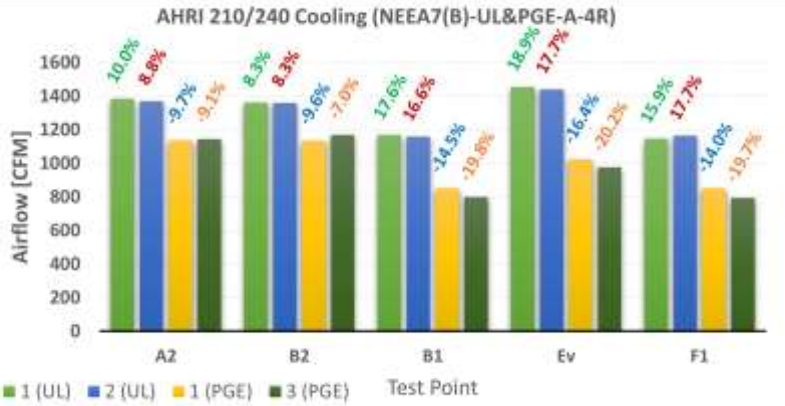


Figure 212. NEEA7(B) (UL&PGE) cooling test intervals airflow rate comparison in different tests

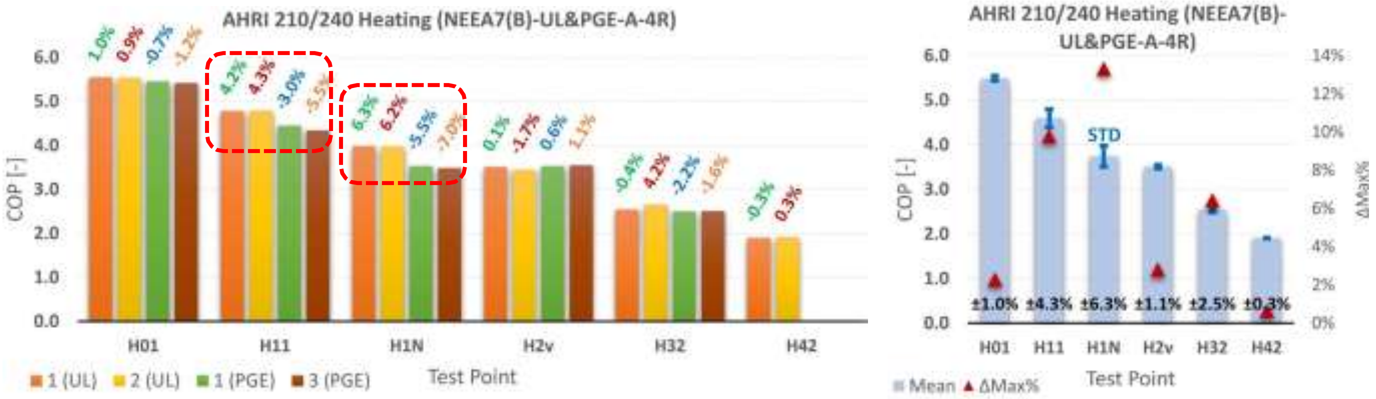


Figure 213. NEEA7(B) (UL&PGE) heating test results a) COP comparison b) Test intervals mean COP with STD and maximum difference for AHRI 210/240 reproducibility assessment

Figure 213 shows the heating test intervals COP comparison, with the maximum difference in COP among four tests varying from 2.3% to 13.3% and STD from ±1% to ±6.3%. It should be noted that at PG&E, test unit performance was not measured in optional test H42. Across all heating test intervals, lower airflow was measured in PG&E tests compared to UL as shown in Figure 214, possibly due to higher external static pressure, similar to cooling results. However, unlike cooling test results, heating rate and power consumption were observed lower, higher, and similar in PG&E heating tests compared to UL at the same test conditions as shown in Figure 215 and Figure 216. In test intervals, H11 and H1N, relatively large differences were noted between two lab test results with higher COPs measured in UL tests compared to PG&E tests. However, there is not a clear pattern in measured airflow and test room conditions that can explain these differences in heating performance between the two labs. So the observed differences are possibly due to variation in test unit setup, its control settings, test facility control, and instrumentation between the two labs.

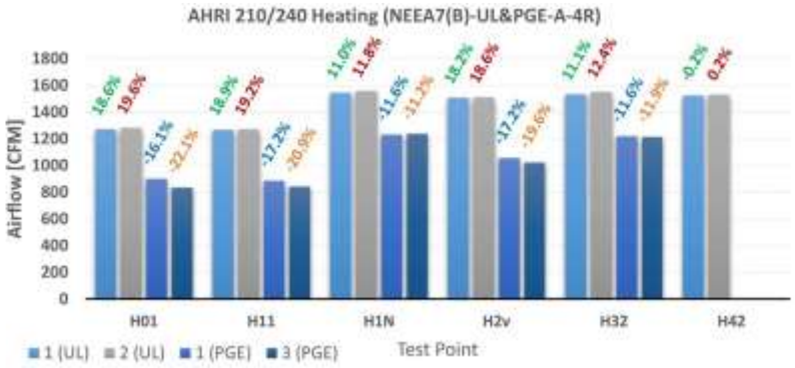


Figure 214. NEEA7(B) (UL&PGE) heating test intervals airflow rate comparison in different tests

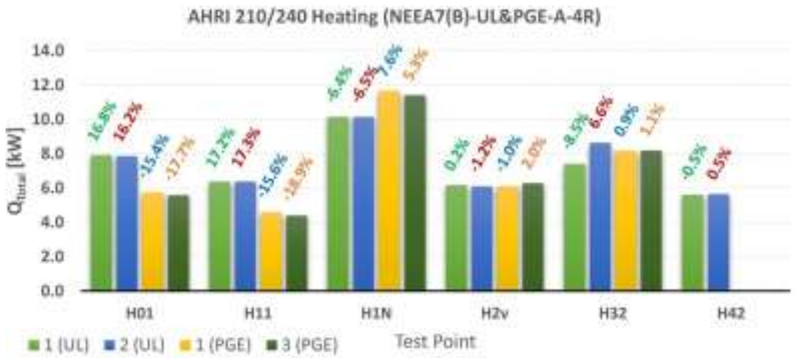


Figure 215. NEEA7(B) (UL&PGE) heating test intervals heating rate comparison in different tests

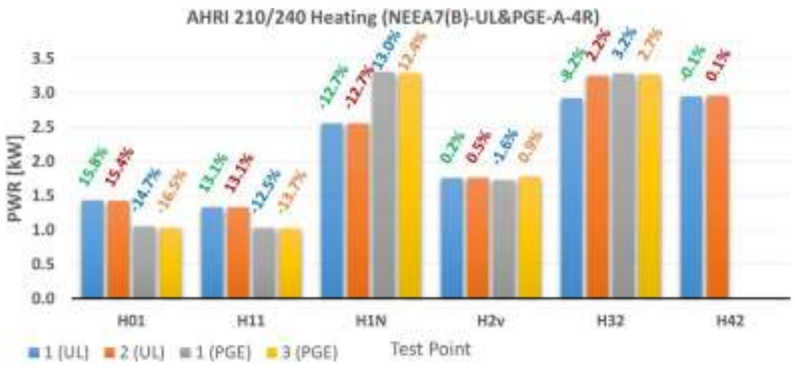


Figure 216. NEEA7(B) (UL&PGE) heating test intervals power consumption rate comparison in different tests

Table 35 shows the comparison in estimated SEER and HSPF based on NEEA7(B) four sets of tests in two labs in dataset NEEA7(B)-UL&PGE-A-4R. Overall good reproducibility was observed in seasonal performance metrics. For cooling, the maximum difference in estimated SEER between different tests was around 3.9%, STD around ±1.6%, and 95% CI around ±2.9%. The lowest SEER was estimated based on test #5 results at PG&E and the maximum was estimated based on test #9 at UL and test #7 at PG&E. Further, the maximum difference in HSPF among 4 tests was around 4.6% with STD around ±1.8%. and 95% CI around ±3.4%.

Table 35. NEEA7(B)-UL&PGE-A-4R dataset AHRI 210/240 seasonal performance reproducibility summary

Metric	7 (UL)	9 (UL)	5 (PGE)	7 (PGE)	∆ [Max - Min] %	95% CI	STD
SEER2	16.03	16.10	15.50	16.10	3.8%	±2.9%	±1.6%
HSPF2	9.11	9.45	9.54	9.50	4.6%	±3.4%	±1.8%

9 AHRI 210/240 Repeatability and Reproducibility Summary

In this section, an overall summary of the AHRI 210/240 repeatability and reproducibility assessment is provided based on the test results of different units in two labs. Figure 217 shows the average SEER and HSPF for each unit based on corresponding datasets as defined in Table 3 from UL and PG&E labs along with a 95% confidence interval (CI) for AHRI 210/240 repeatability and reproducibility evaluation. As 95% CI is based on a *t*-distribution assumption which is a strong function of the number of samples, that might lead to erroneous conclusions if comparing 95% CI for repeatability evaluation with two samples and reproducibility assessment with four samples for the same unit. On the other hand, the standard deviation is a better statistical parameter that is normalized with the number of samples. Table 36 shows the summary of AHRI 210 repeatability and reproductivity assessment with the average value of SEER and HSPF based on different tests along with the percentage difference in maximum and minimum value, STD, and 95% CI as statistical parameters show variability in estimated seasonal performance based on different tests.

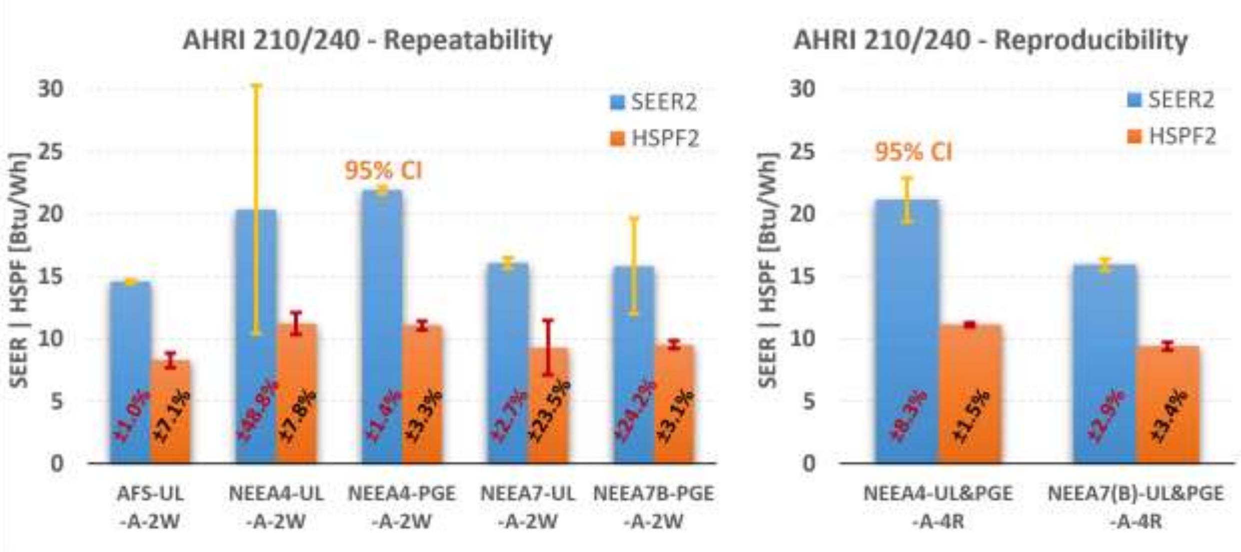


Figure 217. AHRI 210/240 repeatability and reproducibility summary with average seasonal performance metric average and 95% CI based on different tests

Table 36. AHRI 210/240 repeatability and reproducibility assessment summary

AHRI 201/240		Repeatability					Reproducibility	
Metric	**Dataset**	**AFS-UL-A-2W**	**NEEA4-UL-A-2W**	**NEEA4-PGE-A-2W**	**NEEA7-UL-A-2W**	**NEEA7 B-PGE-A-2W**	**NEEA4-UL&PGE-A-4R**	**NEEA7(B)-UL&PGE-A-4R**
SEER2	Mean	14.56	20.34	21.91	16.06	15.80	21.12	15.93
	Δ(Max - Min) [%]	0.2%	7.7%	0.2%	0.4%	3.8%	11.2%	3.8%
	STD [%]	±0.1%	±3.8%	±0.1%	±0.2%	±1.9%	±4.5%	±1.6%
	95% CI [%]	±1.0%	±48.8%	±1.4%	±2.7%	±24.2%	±8.3%	±2.9%
HSPF2	Mean	8.27	11.20	11.06	9.28	9.52	11.13	9.40
	Δ(Max - Min) [%]	1.1%	1.2%	0.5%	3.7%	0.5%	2.2%	4.6%
	STD [%]	±0.6%	±0.6%	±0.3%	±1.8%	±0.2%	±0.8%	±1.8%
	95% CI [%]	±7.1%	±7.8%	±3.3%	±23.5%	±3.1%	±1.5%	±3.4%

AFS unit showed good repeatability at UL in two repeated tests as shown with the AFS-UL-A-2W dataset results. Good repeatability was also observed for the NEEA4 unit at UL and PG&E as shown with datasets NEEA4-UL-A-2W and NEEA4-PGE-A-2W, except for cooling test results at UL in dataset NEEA4-UL-A-2W. In cooling test intervals at UL with NEEA4, relatively poor performance was measured in test #7 in dataset NEEA4-UL-A-2W, resulting in around 7.7% lower SEER compared to test #9 set of tests and STD around ±3.8%. Further, for the NEEA4 unit, higher SEER was estimated based on PG&E tests in dataset NEEA4-PGE-A-2W compared to UL tests in dataset NEEA4-UL-A-2W, resulting in relatively poor reproducibility in cooling test results with STD of around ±4.5% and CI around ±8.3% based on four tests in dataset NEEA4-UL&PGE-A-4R. However, the NEEA4 unit had slightly better HSPF that was estimated based on PG&E tests in dataset NEEA4-PGE-A-2W compared to UL tests in dataset NEEA4-UL-A-2W, and overall good reproducibility was noted in heating test results with STD of ±0.8% and 95% CI of ±3.3% based on four tests in dataset NEEA4-UL&PGE-A-4R. NEEA7 test results showed good repeatability in the UL tests in dataset NEEA7-UL-A-2W as well as NEEA7B unit in PG&E lab tests in dataset NEEA7B-PGE-A-2W. At UL, relatively better repeatability was noted in cooling test results compared to heating in dataset NEEA7-UL-A-2W, whereas the opposite was observed in PG&E test results in dataset NEEA7B-PGE-A-2W. Between the two labs, overall good reproducibility was observed for cooling as well as heating results based on dataset NEEA7(B)-UL&PGE-A-4W with 95% CI of around ± 2.9% in SEER and around ±3.4% in HSPF. The differences in the two labs are possibly due to differences in the test setup, charge, instrumentation, airflow rate measurement (Figure 207 and Figure 212), external static pressure control, and measurement uncertainties.

Table 37 shows the overall AHRI 210/240 repeatability and reproducibility evaluation results with a qualitative flag for different datasets based on different test units in two labs, similar to EXP07 results presented previously.

Table 37. AHRI 210/240 repeatability and reproducibility status summary for different datasets

Dataset	Repeatability Status	
	Cooling	Heating
AFS-UL-A-2W	Good	Good
NEEA4-UL-A-2W	Fair	Good
NEEA7-UL-A-2W	Good	Good
NEEA4-PGE-A-2W	Good	Good
NEEA7B-PGE-A-2W	Good	Good
Reproducibility Status		
NEEA4-UL&PGE-A-4R	Fair	Good
NEEA7(B)-UL&PGE-A-4R	Good	Good

Figure 218 and Figure 219 compares the repeatability and reproducibility of EXP07 and AHRI 210/240 in terms of percentage STD applied to cooling and heating seasonal performance metrics using the test data from UL and PG&E for NEEA4 and NEEA7(B) (NEEA7 & NEEA7B). STD is used as a statistical parameter rather than 95% CI because a different number of tests were used in the AHRI 210/240 and EXP07 assessments for some cases. For EXP07 results, a mean and range of cooling and heating SCOP STD as observed across different climate zones is shown, whereas single SEER and HSPF STD values are presented for AHRI results in a single climate zone. The results for reproducibility will first be discussed, which can be assessed by viewing the results for all tests of each unit across both labs in Figure 218 and Figure 219. The datasets utilized for EXP07 and AHRI 210/240 reproducibility assessment for NEEA4 and NEEA7(B) can be seen in Table 3. For both NEEA4 and NEEA7(B), significantly better reproducibility is observed in AHRI 210/240 results compared to EXP07, except for NEEA7(B) heating results, where comparable variations in seasonal performance metrics were observed for both testing approaches. Some possible reasons for the poor reproducibility of EXP07 have previously been discussed.

Repeatability can be assessed by viewing results in Figure E.6 and Figure E.7 that were determined from test results at each individual lab (UL and PG&E). Note that the AHRI 210/240 repeatability tests for units NEEA4 and NEEA7(B) involved only two tests at each lab that were performed sequentially without having to reinstall the unit as defined in datasets NEEA4-UL-A-2W, NEEA4-PGE-A-2W, NEEA7-UL-A-2W, and NEEA7B-PGE-A-2W. The EXP07 tests at UL involved three tests where only two of the tests were back-to-back without reinstallation as defined in datasets NEEA4-UL-E-3R and NEEA7(B)-UL-E-3R. In order to provide fairer repeatability comparisons, results for UL EXP07 that only include back-to-back tests without reinstallation are also shown (NEEA4-UL-E-2W and NEEA7-UL-E-2W datasets). Overall, the repeatability of both EXP07 and AHRI 210/240 are good and comparable in both cooling and heating when not considering data where the unit was not reinstalled. In fact, there are some examples where the EXP07 repeatability was better than for AHRI 210/240 (e.g., cooling tests for NEEA4 at UL). If data is included for units that were reinstalled at UL, then the repeatability of AHRI 210/240 would generally be better than that for EXP07, especially for the NEEA4 heating results. However, it is felt that these comparisons are less relevant for repeatability because they contain several factors related to reproducibility. The differences in the repeatability and reproducibility between the two testing approaches are mainly due to the dynamic nature of tests as per EXP07 with a test unit embedded with controls and thermostat, whereas steady-state tests are mainly performed based on AHRI 210/240 at fixed compressor speeds and airflows. The other factors which contribute to these variations in the two testing approaches are the differences in temperature bin method and bin hours fraction.

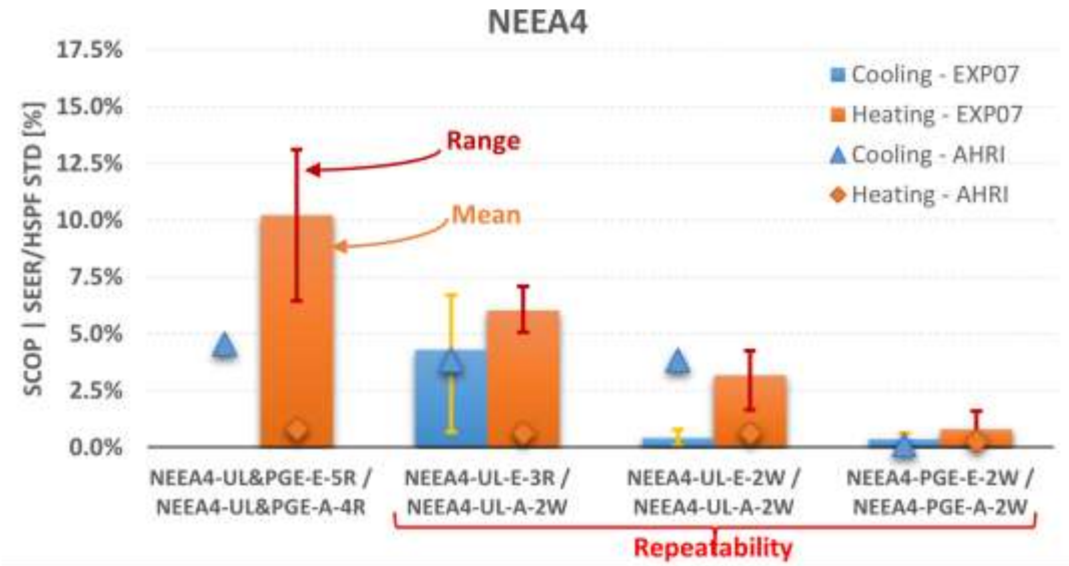


Figure 218. NEEA4 EXP07 and AHRI 210/240 seasonal performance metrics STD comparison for reproducibility and repeatability assessment. EXP07 SCOP STD mean and range across different climate zones and AHRI SEER / HSPF STD for single climate zone

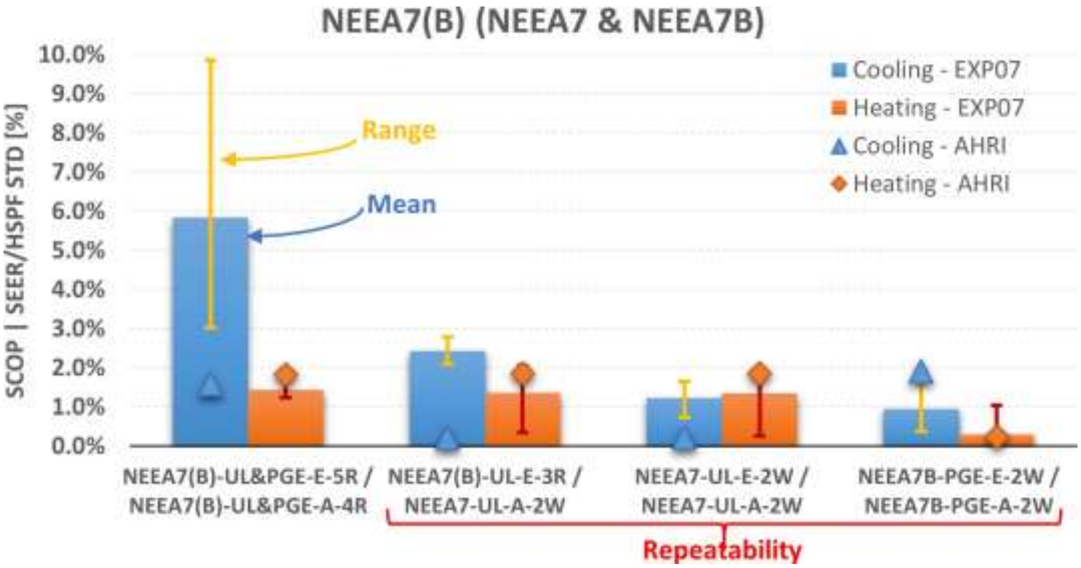


Figure 219. NEEA7(B) EXP07 and AHRI 210/240 seasonal performance metrics STD comparison for reproducibility and repeatability assessment. EXP07 SCOP STD mean and range across different climate zones and AHRI SEER / HSPF STD for single climate zone

10 Conclusions

In this report repeatability and reproducibility assessment of two testing and rating methodologies for residential unitary equipment is presented. One is based on a newly developed load-based testing approach as per CSA EXP07:19, whereas the other is based on the current testing and rating approach as per AHRI 210/240-2023. Round robin tests data of five different heat pumps tested across two different labs, UL and PG&E, was analyzed to perform a quantitative and qualitative comparison of measured performance in different tests at the same test conditions. For each test unit, performance measured in each test interval as well overall estimated seasonal performance metrics were compared for repeatability evaluation in each lab and overall reproducibility across two labs for both test methodologies. Table 38 shows CSA EXP07:19 and AHRI 210/240-2023 overall repeatability and reproducibility results for different units in two labs with a qualitative status. Overall reasonable to good repeatability was observed in EXP07 results in both labs if only comparing the repeated tests without test unit re-installation in between. Considering all the three repeated tests at UL in repeatability evaluation, which is actually a somewhat hybrid between true repeatability and reproducibility analysis, poor repeatability was observed with NEEA4-UL-E-3R dataset, and reasonable repeatability was noted for NEEA7(B)-UL-E-3R dataset. For EXP07 reproducibility evaluation between two labs, NEEA4 cooling results were inconclusive due to a testing approach issue and poor reproducibility was observed in heating test results for NEEA4-UL&PGE-E-5R dataset. NEEA7(B) showed good reproducibility in EXP07 cooling test results, however, comparatively poor reproducibility was observed in cooling tests for NEEA7(B)-UL&PGE-E-5R dataset. On the other hand, as expected, AHRI 210/240 test results for both units showed overall good repeatability as well as reproducibility.

Table 38. EXP07 and AHRI 210/240 repeatability and reproducibility status for different datasets from two labs UL and PG&E

CSA EXP07:2019			AHRI 210/240-2023		
Dataset	Repeatability		Dataset	Repeatability	
	Cooling	Heating		Cooling	Heating
AFS-UL-E-3R	Good	Good	AFS-UL-A-2W	Good	Good
NEEA4-UL-E-3R	Poor	Poor	-		
NEEA4-UL-E-2W	Good	Fair	NEEA4-UL-A-2W	Fair	Good
NEEA7(B)-UL-E-3R	Fair	Good	-		
NEEA7-UL-E-2W	Good	Good	NEEA7-UL-A-2W	Good	Good
NRCan8-UL-E-2W	Good	Good	-		
NRCan10-UL-E-2W	Good	Good	-		
NEEA4-PGE-E-2W	Good	Good	NEEA4-PGE-A-2W	Good	Good
NEEA7B-PGE-E-2W	Good	Good	NEEA7B-PGE-A-2W	Good	Good
Reproducibility Status					
NEEA4-UL&PGE-E-5R	-	Poor	NEEA4-UL&PGE-A-4R	Fair	Good
NEEA7(B)-UL&PGE-E-5R	Fair	Good	NEEA7(B)-UL&PGE-A-4R	Good	Good

Further, a root cause analysis of the observed performance variations was conducted, and possible causes for observed differences were highlighted. In addition, some areas of improvement for the

EXP07 load-based testing methodology were also identified. Some of the important factors that affect the reproducibility include issues related to the installation and setup of the unit for testing (e.g., instrumentation, refrigerant charge, duct static pressure, etc.), different interpretations and implementations of the draft standard, and different characteristics of the environmental chambers. Although these factors are also relevant for the current standard (AHRI 210/240), their impact on the results for load-based testing might be more significant because the draft standard is new and more complicated and the dynamic behavior of the unit with its integrated controls may be sensitive to some installation effects. In addition to the effect of installation and implementation issues on test unit dynamic behavior, variations in test unit dynamic response at the same test conditions could be due to adaptive/learning behavior of the embedded controller and the limited time available for testing. Some of the differences in implementation will likely be resolved as the standard matures and personnel become more familiar with its application. Some of the issues with differences in facilities could be addressed by facility and test equipment improvements. For example, utilizing a thermostat apparatus to provide a standardized environment for the thermostat could reduce differences that are caused by non-uniform airflow and temperatures within environmental chambers. However, differences related to changes in test unit controller behavior would be challenging to address with the test standard without dramatically increasing the testing time. In fact, one of the merits of the load-based testing approach is that it can capture the impacts and sensitivities of test unit performance to dynamic responses of its embedded controller. If a test unit controller performs inconsistently with load-based testing in the laboratory, then it is likely to perform similarly in the field and it is important to capture these effects in a standard method of test. This could be an incentive for manufacturers to develop controllers with more consistent and predictable behavior. With that being said, it would be a good idea to further investigate how best to test and rate units that have inconsistent controller behavior using load-based testing in a repeatable and representative fashion for making the standard more inclusive.

There were likely some differences in implementation of EXP07 convergence criteria that led to some significant performance differences between tests carried out in different labs. For instance, differences in test unit operation mode and dynamic behavior that were captured during periods that were deemed to be converged were likely due to different interpretations of the EXP07 convergence criteria and some limitations of the current convergence criteria. These convergence criteria issues occurred when there was a combination of different operating modes that occurred at a given test condition, such as a sequential combination of unit on/off cycling and defrost, or when there was an irregular on/off cycling pattern with a mixture of short and regular cycles. These issues are discussed in detail in the report with recommended improvements which could be used as a reference to update the EXP07 convergence criteria. In addition to adding clarifications to EXP07 to avoid ambiguity, some training material for lab personnel covering the testing approach as well as convergence criteria with some practical examples could be useful.

Another specific issue identified through testing was that indoor temperatures could converge to significantly different values, most notably when comparing test results between labs. This could primarily be related to differences in the thermostat offset settings at the start of testing that are part of implementation of the EXP07 setup procedures Another EXP07 implementation issue that led to disparities between laboratories is related to interpretation of the decision of when to switch to full-load testing. Prematurely switching to full-load testing can have a significant impact on results at relatively high-load conditions that are important in determining seasonal efficiencies.

Some of the possible sources of differences in test unit performance that are related to test setup (e.g., charge, thermal mass, instrumentation location, sensor dynamic responses, air flow measurement, test unit fan settings) and test facility control (e.g., room conditions, external static pressure) could be addressed by improving test setup requirements and/or corrections for the components that have a significant effect on dynamic performance measurements, such as code-tester thermal mass, instrumentation location, sensor response, etc. Also, more appropriate test condition and operating tolerances should be defined to limit the effect of variations in test facility control on overall performance measurement. Both of these enhancements, however, will necessitate additional research and examination because most of the currently used test methods are for steady-state tests rather than dynamic load-based tests. Also, issues with external static pressure control, airflow measurement, and test fan settings that led to differences in the two labs should be further investigated with the goal of updating the current EXP07 testing approach to have better reproducibility.

This report presents several suggestions for improving the draft EXP07:19 standard that resulted from analysis of the repeatability and reproducibility data along with interactions with the project participants and stakeholders. These are mostly related to convergence criteria, test unit setup, transition to full-load testing, unit control settings, airflow measurement, thermostat offsets, and tolerances associated with dynamic measurements. Many of the improvements may have already been addressed in the most recent updates to EXP07:19 that have led to the latest version (EXP07:22) that is scheduled to be published in August 2022. For future work in the near term, it is suggested that the raw test data that was obtained in this project be re-analyzed to assess whether revisions to the convergence criteria address some of the noted issues and improve overall repeatability and reproducibility. For the longer term, it is strongly recommended that a similar study be carried to assess the overall impact of updates that are incorporated in EXP07:22. A second study will benefit from experiences gained during the initial study. However, a second study should include more test units across more than just two test laboratories and be focused on reproducibility. Although there was sufficient data in the initial study to gain an understanding of the repeatability of the EXP07 testing, the data were not sufficient to draw clear conclusions on its reproducibility. Ideally, reproducibility of test results for both EXP07 and AHRI 210/240 should be considered for all test units across all test labs included in a follow up project.

A key question that needs to be addressed in future work is: what is an acceptable tolerance in seasonal performance metric repeatability and reproducibility for load-based testing per EXP07? An initial step that is necessary to answer this question involves carrying out a detailed uncertainty analysis for measured performance that is due to dynamic measurements that are subject to both instrumentation uncertainties and dynamic behavior associated with the dynamic testing approach, the testing facilities, and operating tolerances. This should lead to the definition of minimum tolerances. However, the actual tolerances would need to be larger to account for human factors that lead to differences in test installation and setup in different labs. These tolerances can only be defined through testing data obtained across multiple labs when applying the updated EXP07:22 draft standard. Given the dynamic nature of tests in the load-based testing approach, these tolerances will most likely be larger than what is expected for AHRI 210/240. So, there is undoubtedly a tradeoff when comparing AHRI 210/240 and EXP07 between reproducibility and field representativeness of the test approaches. It is too soon to fully understand this tradeoff before additional work is carried out. In particular, it is important to analyze AHRI 210/240 and EXP07 test results with a goal of identifying the primary sources of differences between the two approaches along with their contributions to overall estimated seasonal performance differences. It is also important to consider paths towards even better testing and rating approaches for the future. In particular, there could be tremendous value in defining test approaches that could determine performance maps for heat pumps and air conditioners that could be integrated into simulation tools for estimating more accurate and climate-specific seasonal performance ratings. Testing of equipment with integrated controls could be important in developing performance maps for this purpose. Future work to develop procedures for testing and performance mapping that minimize test requirements while enhancing reproducibility should be a priority.

References

AHRI. (2017). *AHRI Standard 210/240. Performance Rating of Unitary Air-Conditioning and Air-Source Heat Pump Equipment*. Air-Conditioning, Heating, and Refrigeration Institute.

AHRI. (2020). *AHRI Standard 210/240 - 2023 Performance Rating of Unitary Air-Conditioning & Air-Source Heat Pump Equipment*.

CSA. (2019). *CSA EXP07:19 Load-based and climate-specific testing and rating procedures for heat pumps and air conditioners*.

Appendix A – EXP07 Cooling Test Plots

In this appendix, three test equipment performance, indoor temperature, and relative humidity variations plots are presented for cooling load-based tests based on CSA EXP07:19. The main purpose of this is to provide a reference to the reader for the dynamic variation of indoor temperature and humidity during cooling load-based testing. In dry-coil tests, indoor humidity was kept low such that there is no dehumidification at the indoor unit cooling coil; whereas, in humid coil tests indoor humidity was varied based on the virtual building latent load model and test unit latent cooling rate. In the following plots, a legend similar to the plots shown above is utilized with the addition of indoor dew point temperature ($IISS\,SSPP$) and indoor relative humidity ($IISS\,RRRR$).

AFS

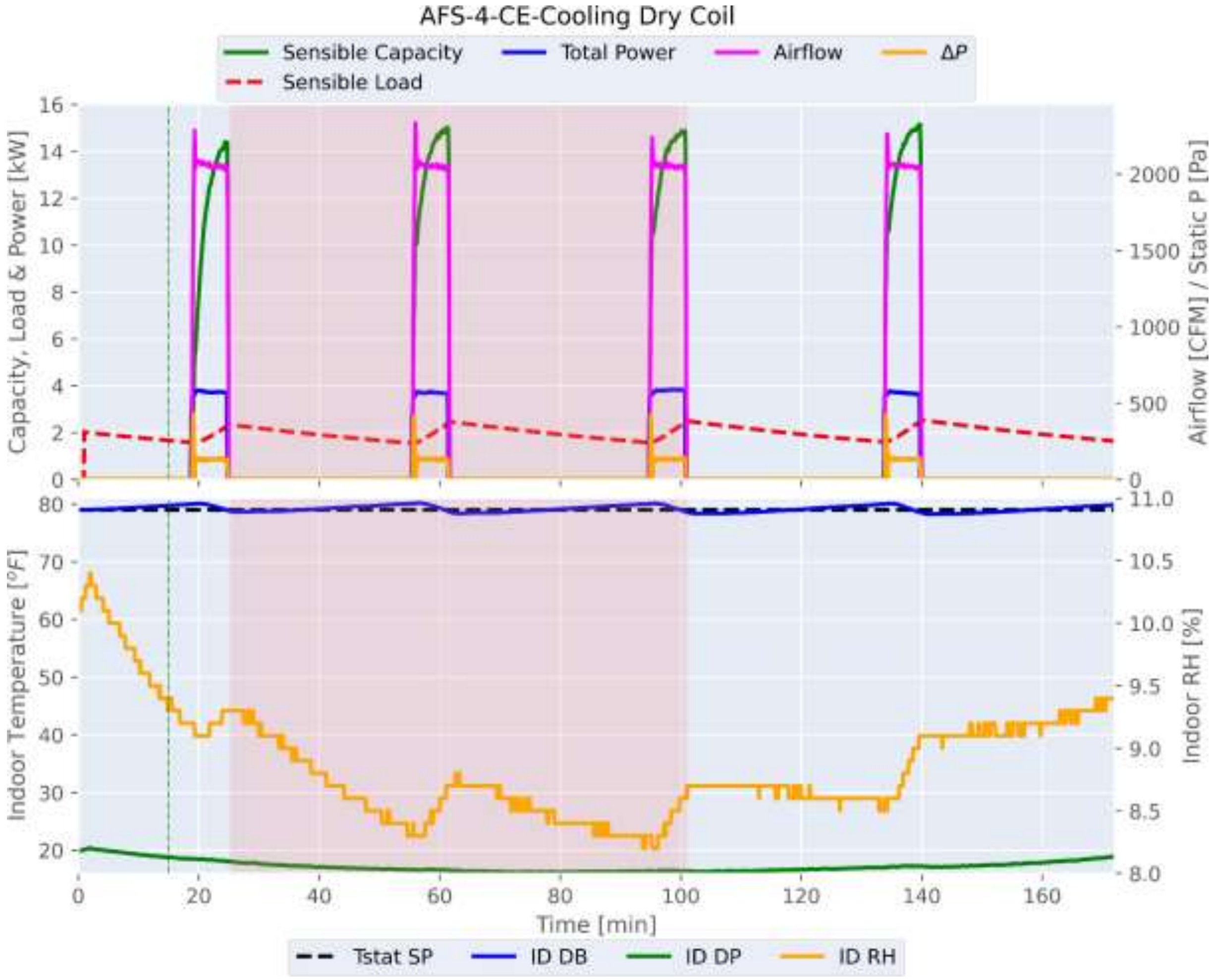


Figure A.1. AFS-4 cooling dry-coil test interval CE (77℉)

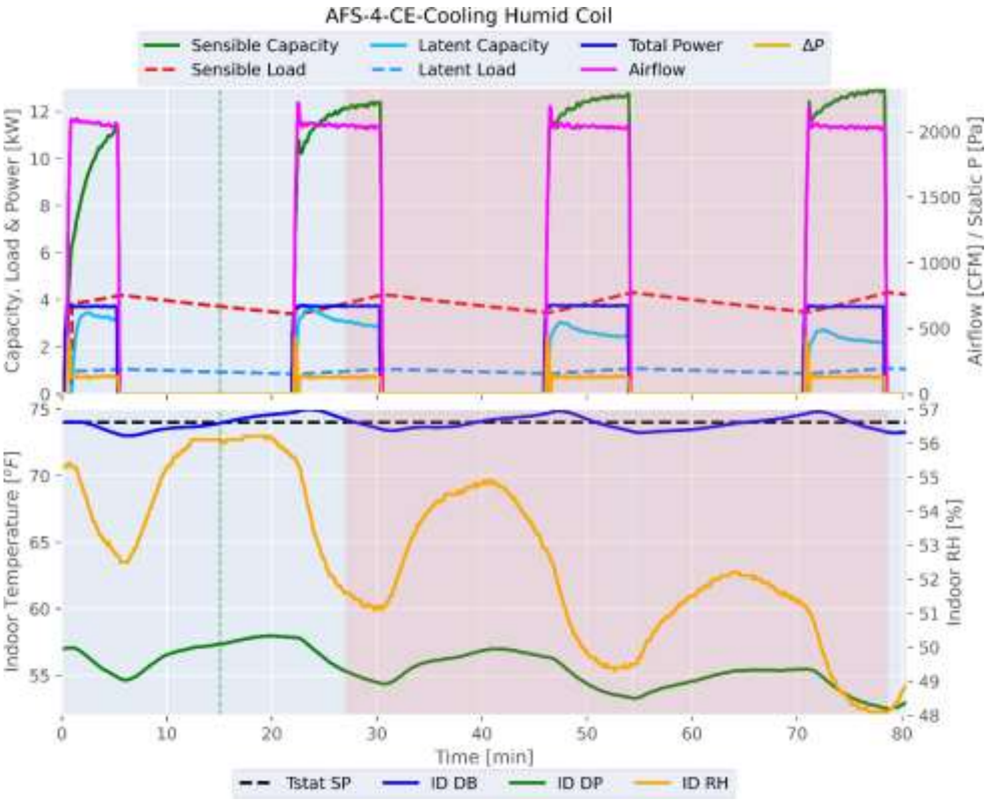


Figure A.2. AFS-4 cooling humid-coil test interval CE (77°F)

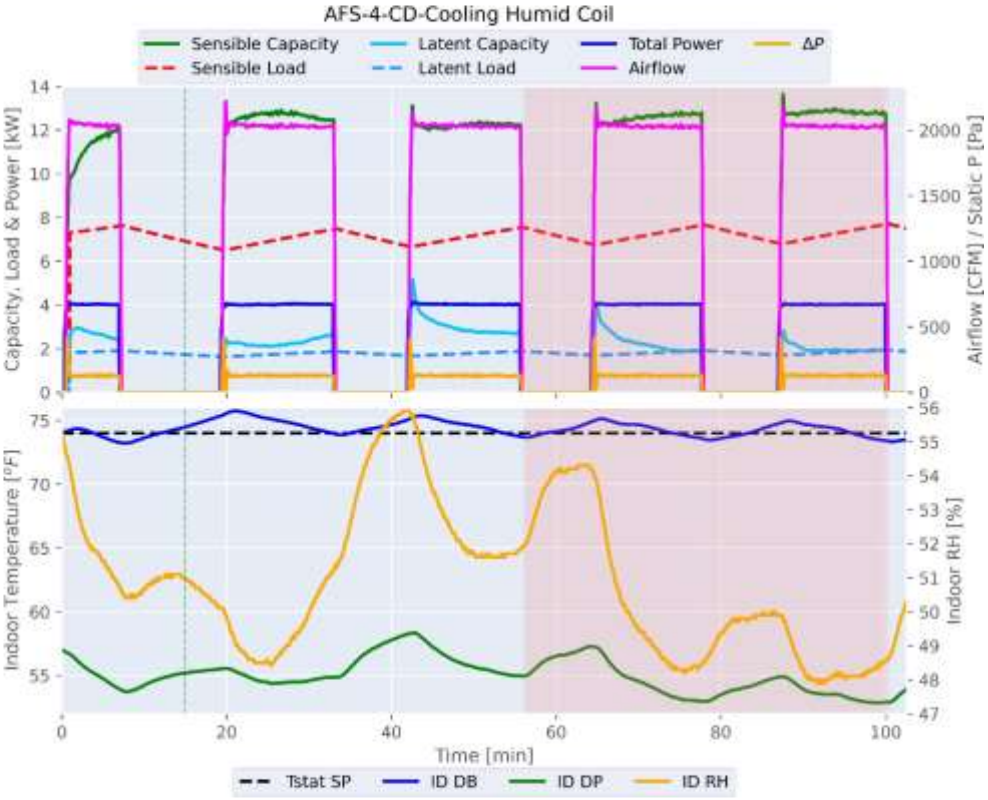


Figure A.3. AFS-4 cooling humid-coil test interval CD (86°F)

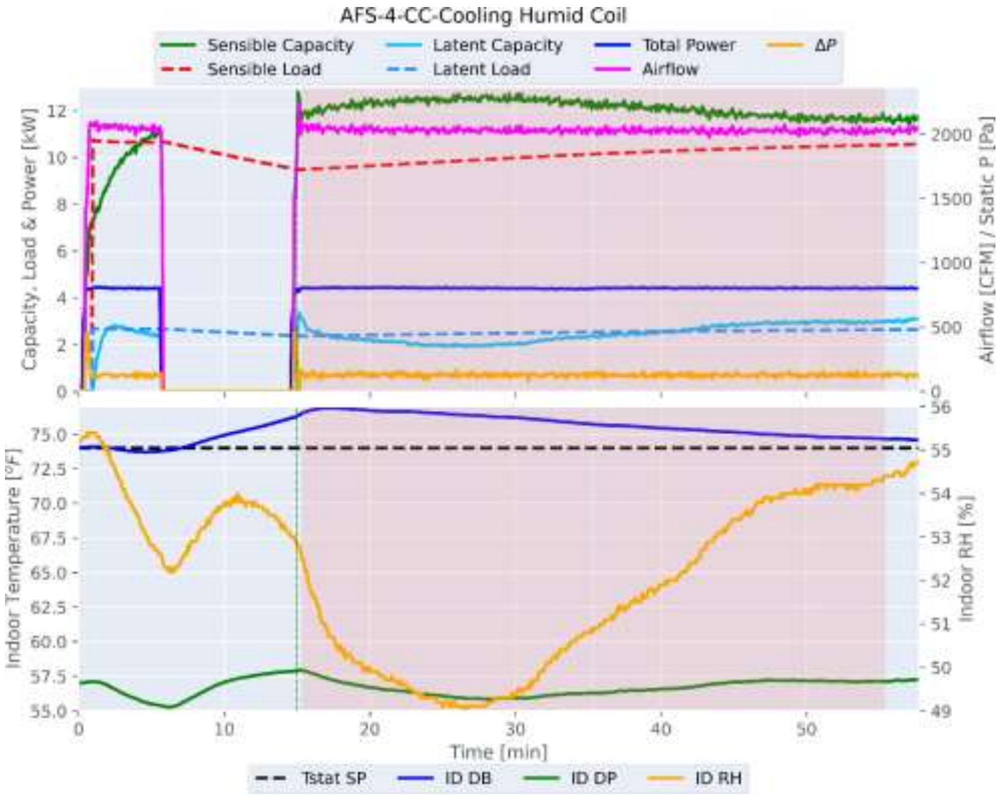


Figure A.4. AFS-4 cooling humid-coil test interval CC (95℉)

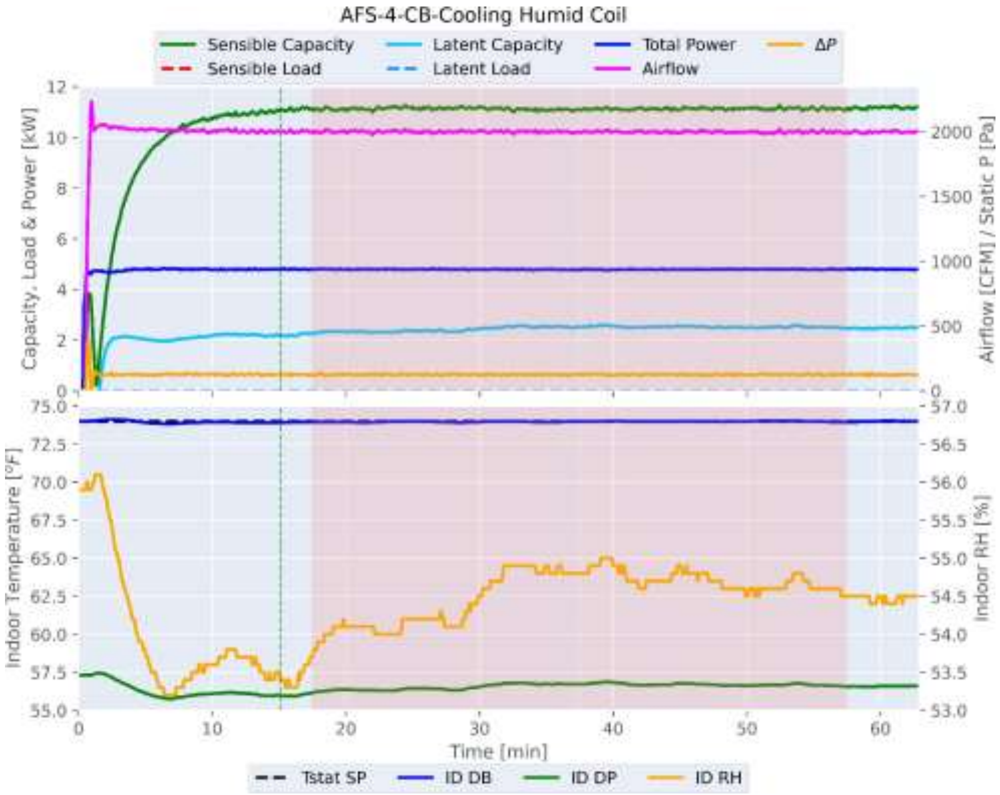


Figure A.5. AFS-4 cooling humid-coil test interval CB (104℉)

NEEA4

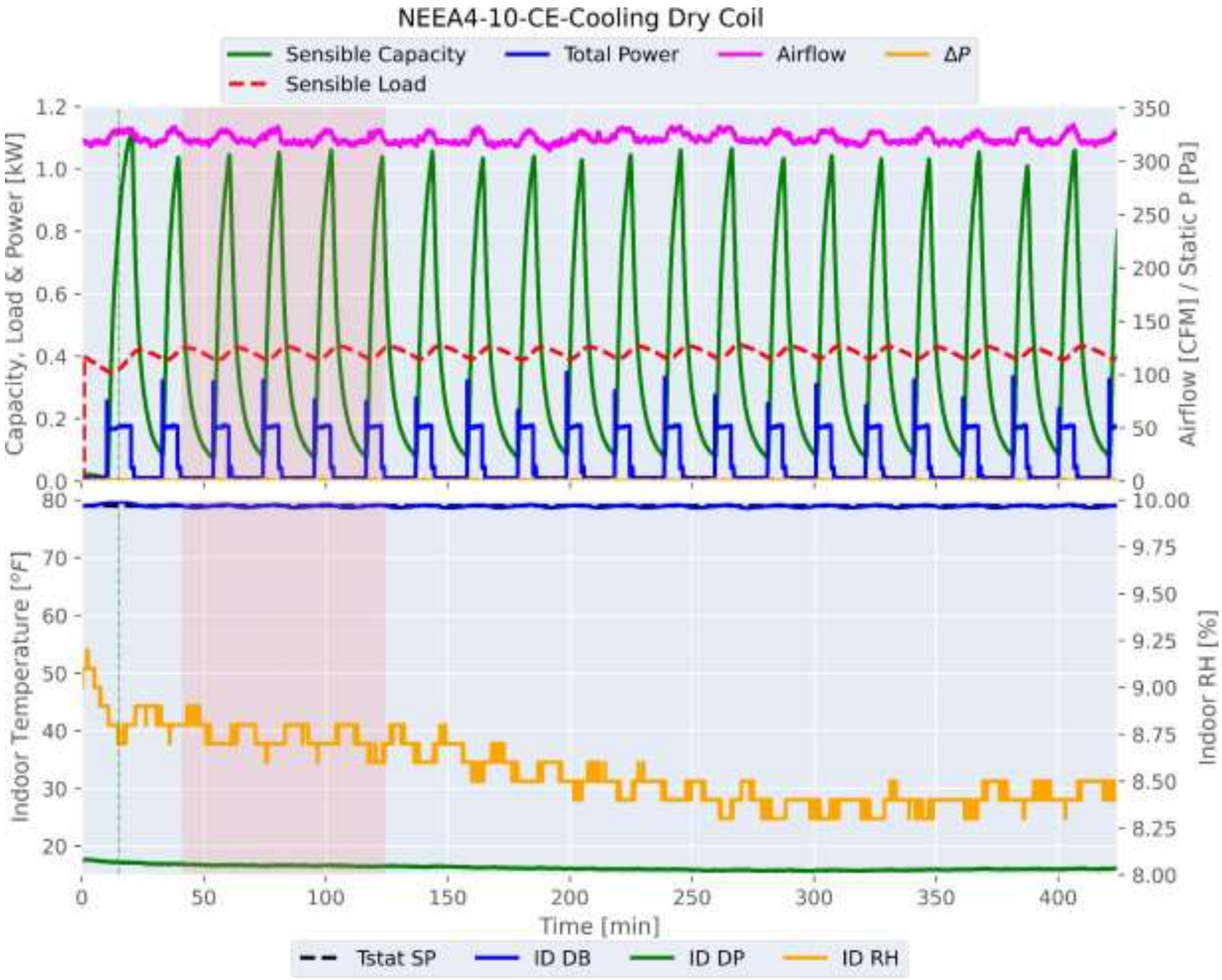


Figure A.6. NEEA4-10 cooling dry-coil test interval CE (77℉)

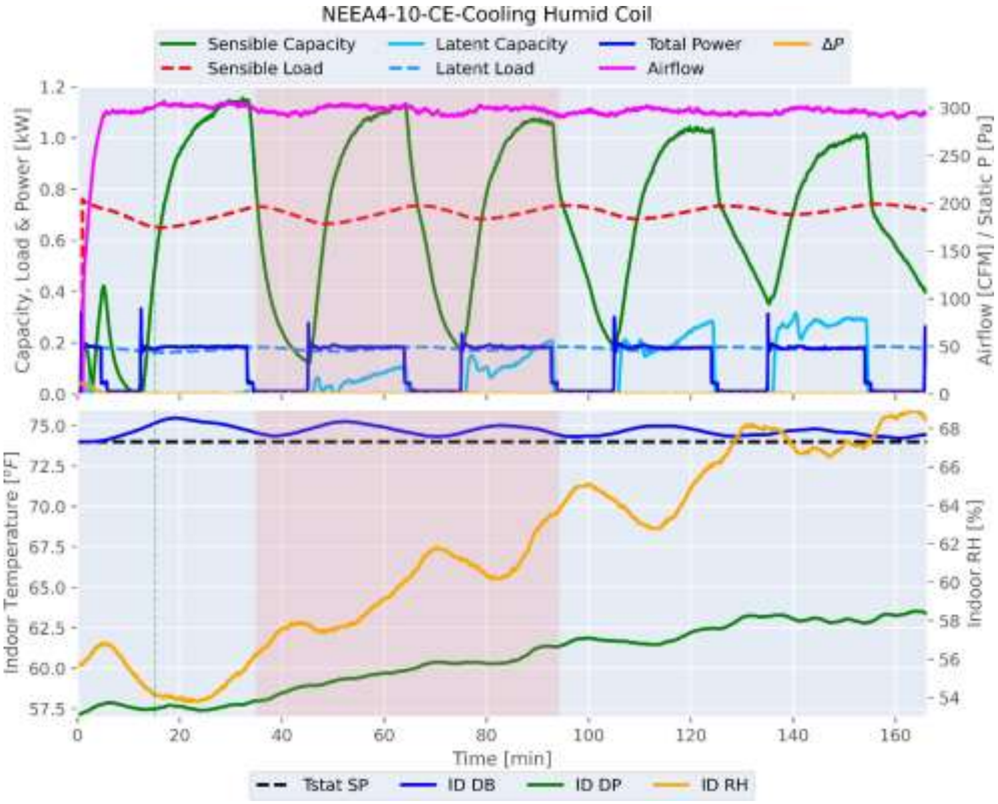


Figure A.7. NEEA4-10 cooling humid-coil test interval CE (77°F)

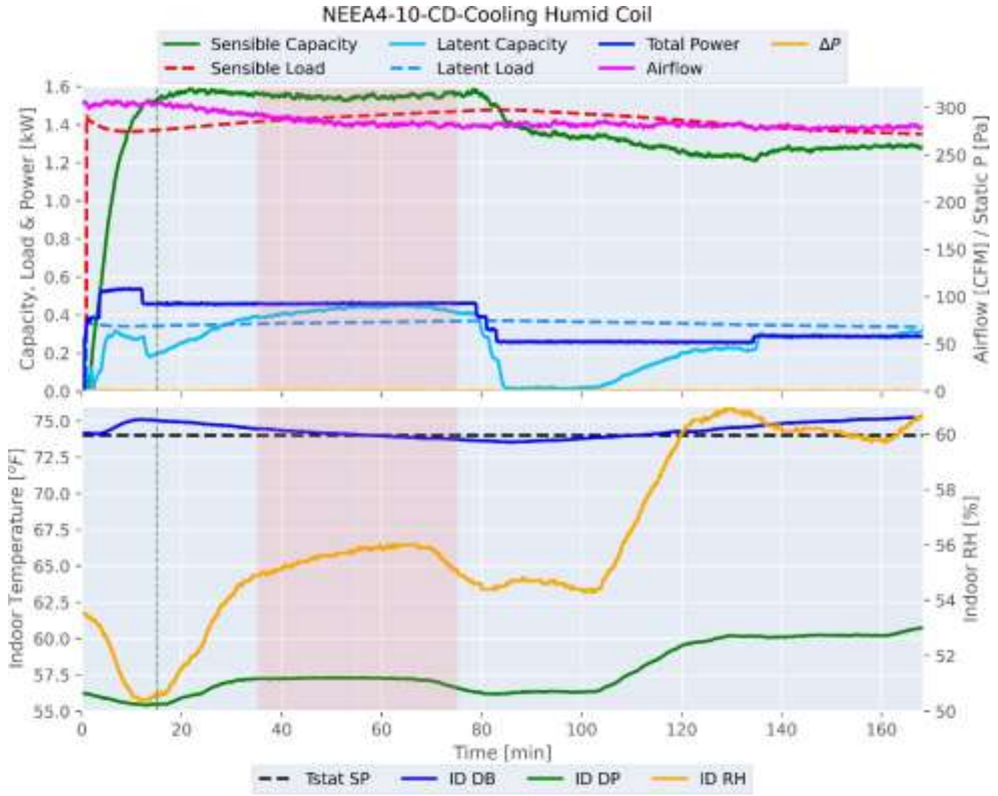


Figure A.8. NEEA4-10 cooling humid-coil test interval CD (86°F)

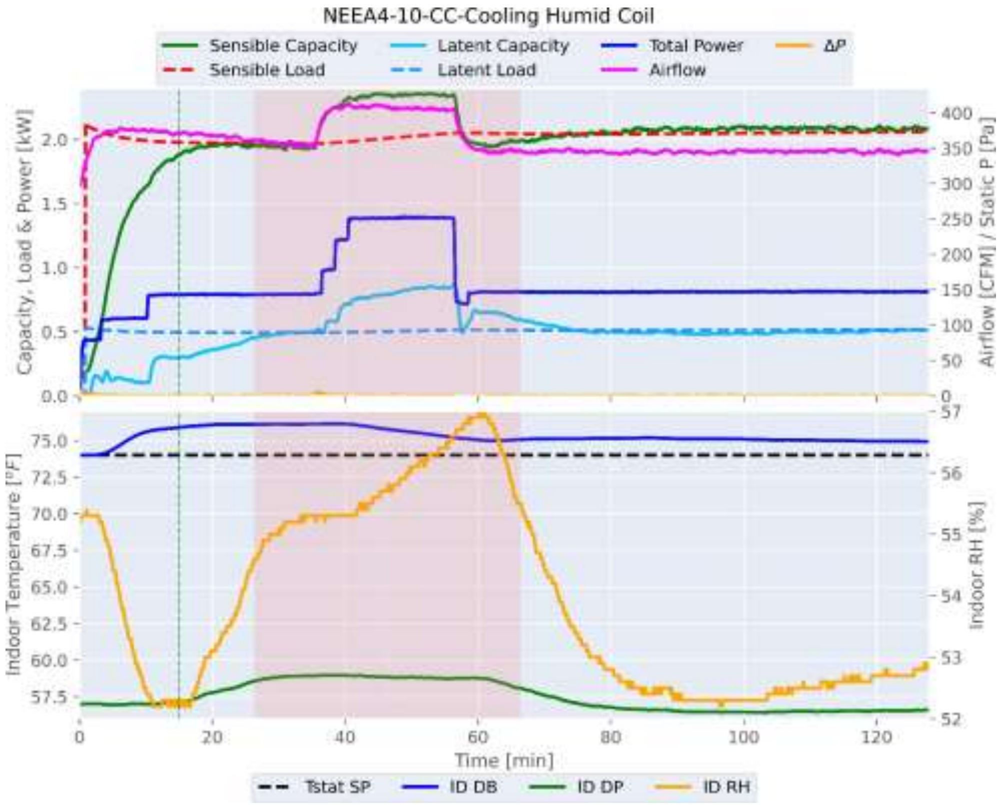


Figure A.9. NEEA4-10 cooling humid-coil test interval CC (95℉)

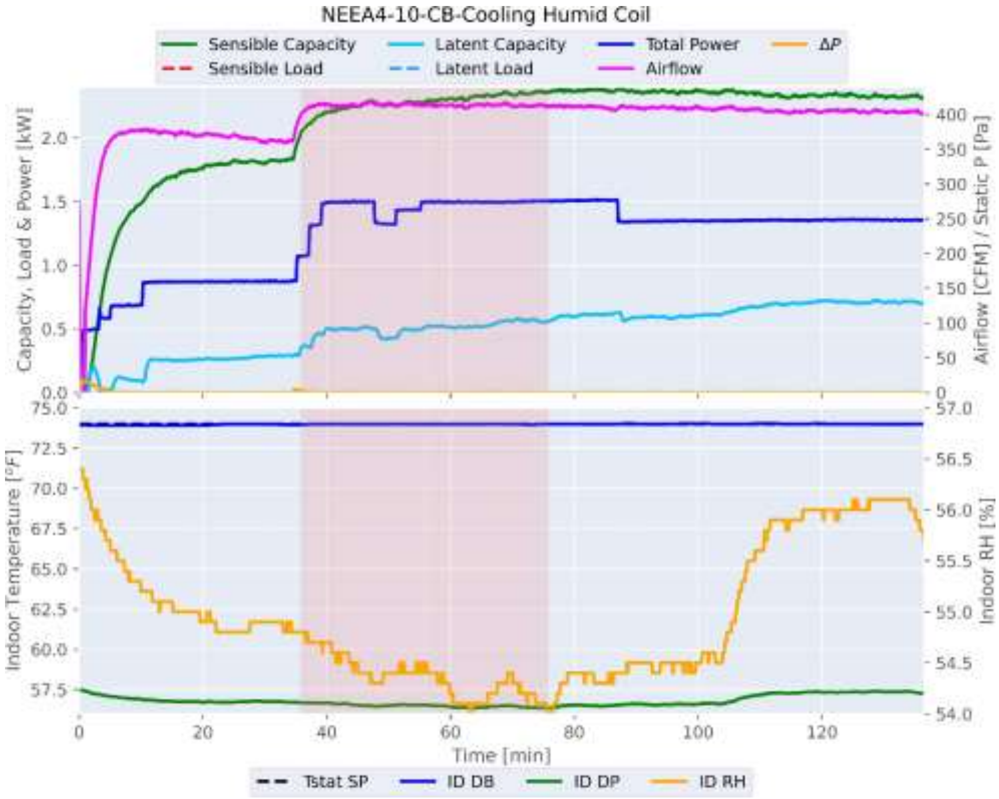


Figure A.10. NEEA4-10 cooling humid-coil test interval CB (104℉)

NEEA7

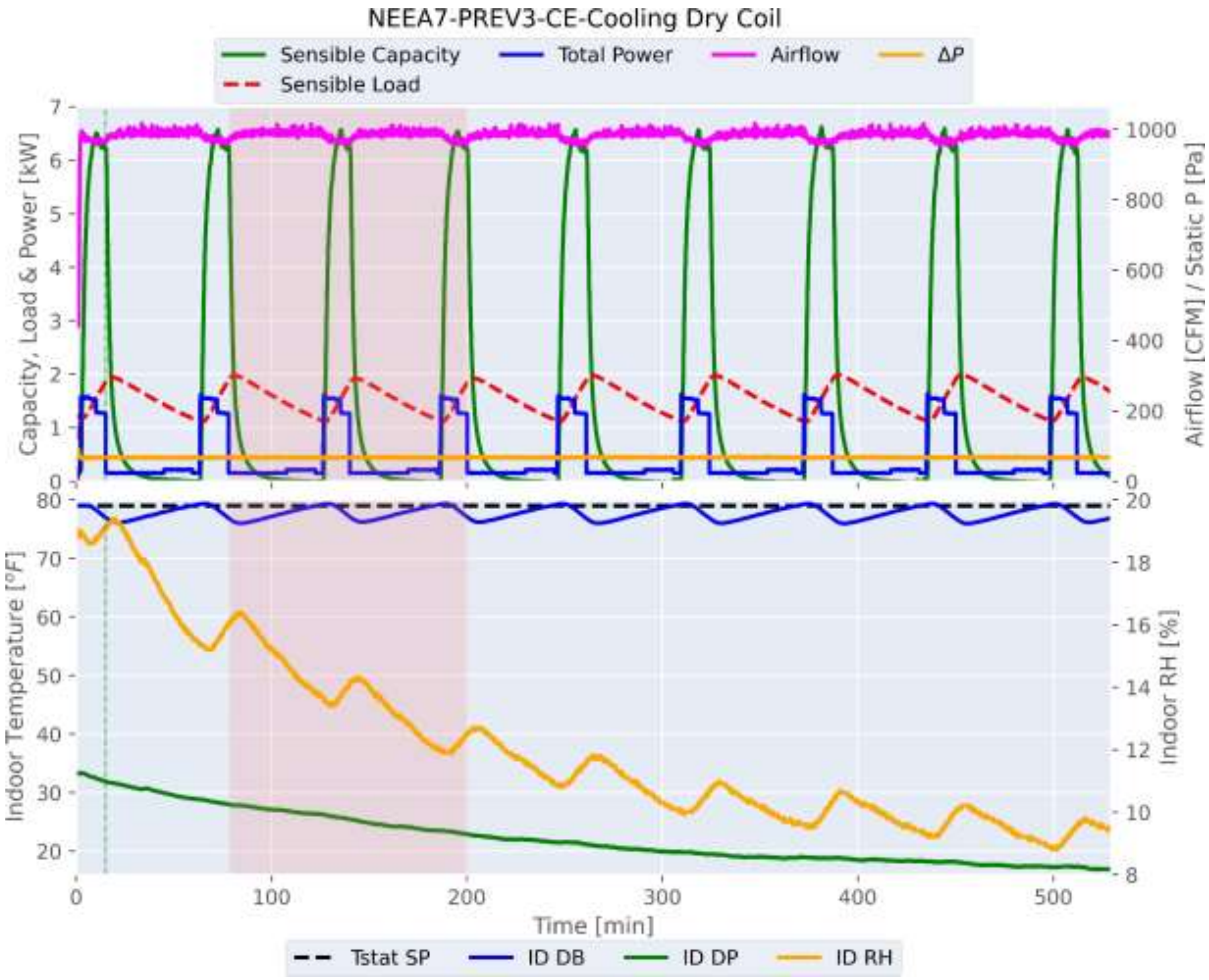


Figure A.11. NEEA7-PREV3 cooling dry-coil test interval CE (77℉)

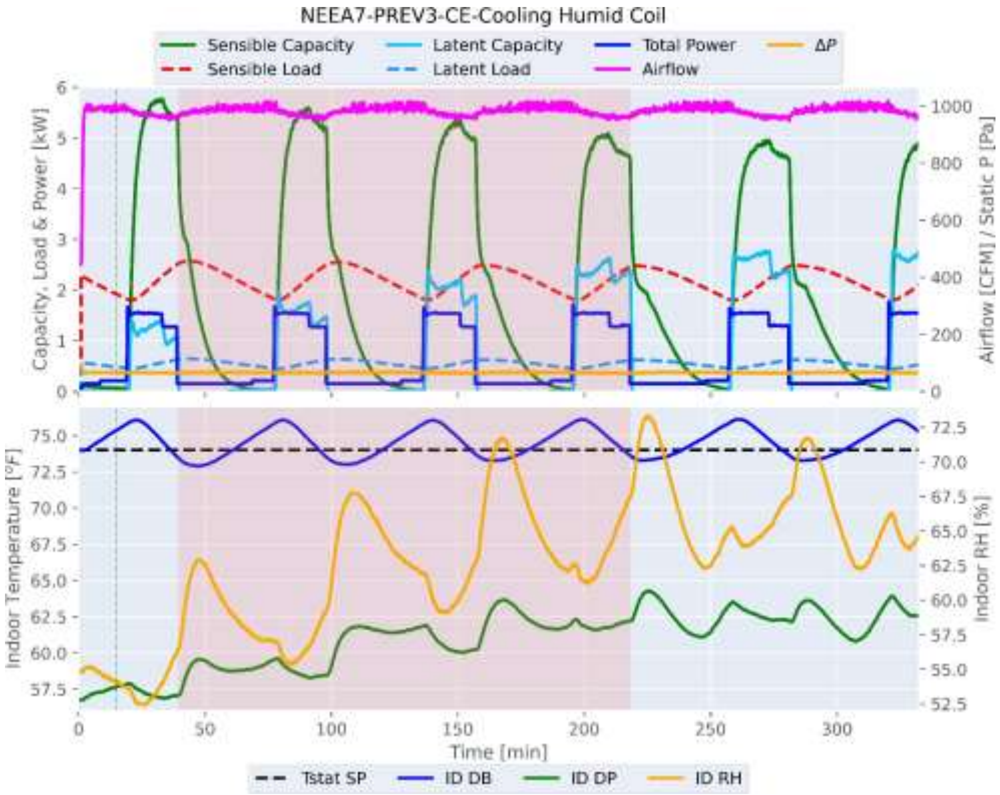


Figure A.12. NEEA7-PREV3 cooling humid-coil test interval CE (77℉)

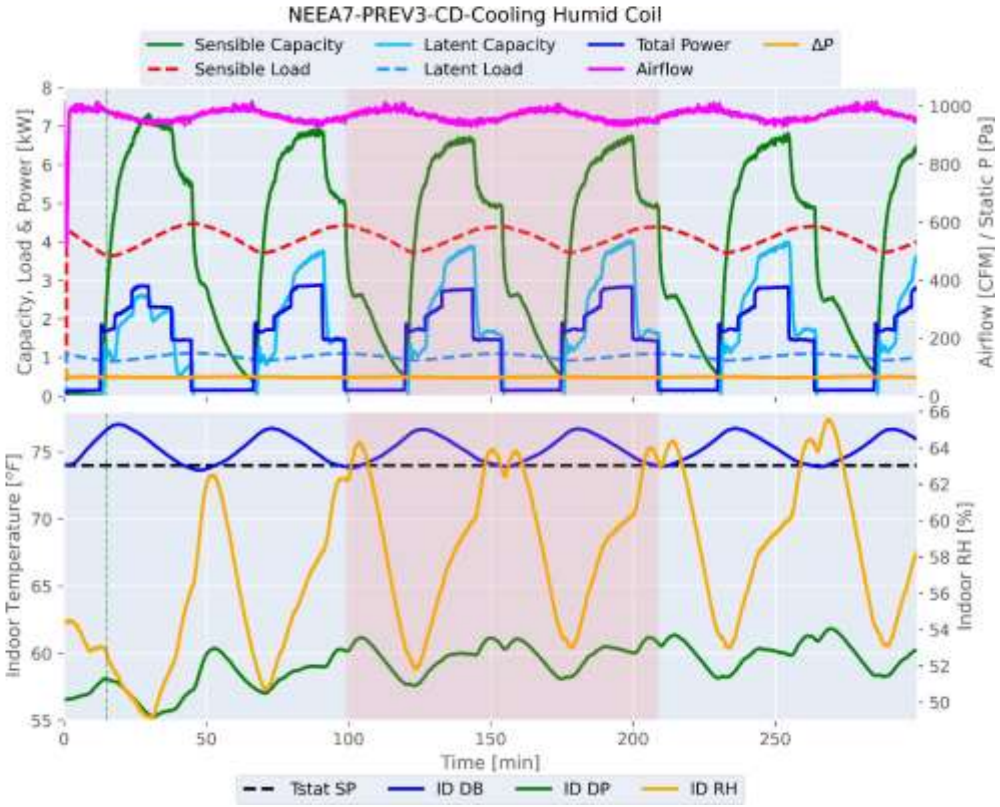


Figure A.13. NEEA7-PREV3 cooling humid-coil test interval CD (86℉)

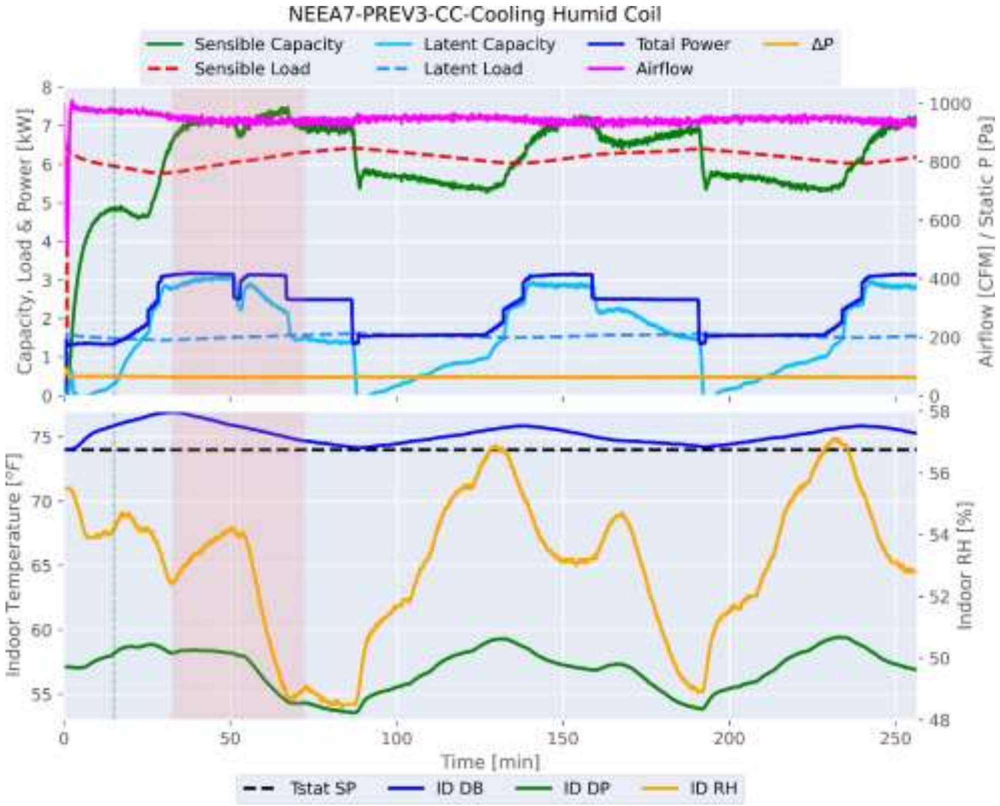


Figure A.14. NEEA7-PREV3 cooling humid-coil test interval CC (95℉)

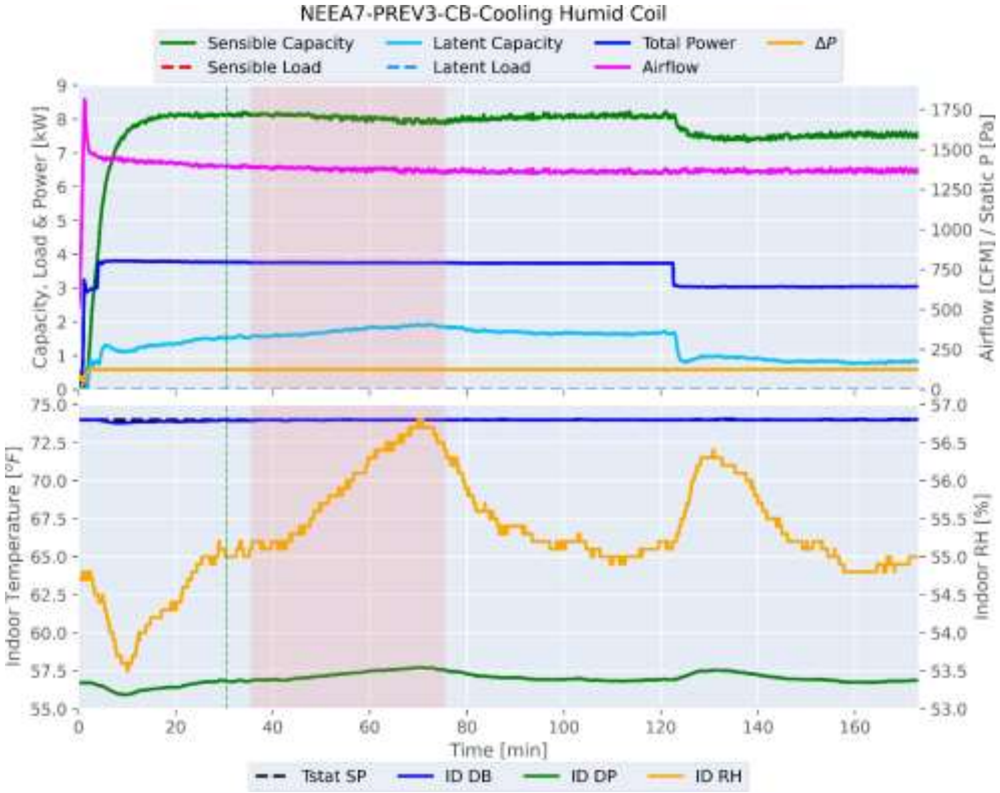


Figure A.15. NEEA7-PREV3 cooling humid-coil test interval CB (104℉)