Mode-Aware Short-Horizon Forecasting for Residential Heat Pump Water Heaters under Realistic Draws

Ali Bohluli Oskouei, Jialin Liu, Sara Eftekharnejad
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY, USA
{abohluli, jliu359, seftekha}@syr.edu

***Abstract*—Heat pump water heaters (HPWHs) are one of the most promising residential loads for shifting demand, yet their highly mode-dependent behavior makes control and forecasting nontrivial. This paper studies 120 V plug-in hybrid HPWHs as flexible electric loads and develops short-horizon, mode-aware power forecasting to support predictive control and demand response. Experiments at the Syracuse Center of Excellence instrumented A.O. Smith hybrid units with second-level telemetry under dynamic draw profiles representative of a three to four-person household. The resulting dataset includes ambient, inlet, and outlet temperatures, operating modes, and power. We benchmark a diverse set of models, from gradient-boosted trees to deep sequence learners. A graph attention network achieved a mean absolute error of 29.84 W, and a model-agnostic meta-learning approach achieved 31.54 W, followed by a feature-token transformer and XGBoost. Long short-term memory, gated recurrent unit, and one-dimensional convolutional networks were competitive in steady regimes but underperformed on rare spikes. Mode choice had a clear impact on daily energy: heat-pump-dominated days consumed 0.8–0.95 kWh, compared with 1.6–1.8 kWh on electric-resistance-dominated days. These results show that accurate, mode-aware forecasting can enable demand-response strategies that preheat off-peak and suppress on-peak operation while maintaining service quality, thereby reducing peak impacts and increasing flexibility. The study provides a high-resolution, data-driven foundation for grid-interactive HPWH control and outlines practical considerations for real-world deployment.**

***Index Terms*—Heat pump water heater, demand response, predictive control, short-horizon forecasting, time-series modeling.**

I. Introduction

This study examines the operational performance and load-shifting behavior of a residential heat pump water heater (HPWH)—the A.O. Smith 120 V plug-in hybrid electric model—under dynamic usage profiles and grid-interactive strategies to assess its contribution to demand-side management and flexible grid integration. The objective is to quantify energy use, thermal performance, and operational response across realistic daily draws using key performance indicators such as power consumption, outlet temperature, and recovery time under multiple modes and setpoints. To this end, a simulation framework ingests setpoint, inlet temperature, and ambient temperature, executes a predefined draw schedule at 1-second resolution, and tracks outlet temperature, internal tank states, and heater activation; peak and off-peak behavior are analyzed to gauge suitability for intelligent load-control programs. The resulting multi-setpoint, multi-mode dataset supports the development of short-horizon power forecasters and enables comparative evaluation across operating modes and temporal segments.

A. Experiment Unit Specifications

Two A.O. Smith 66-gal HPWHs with up-to-140 °F setpoints were evaluated; they differ in efficiency, recovery, and installation requirements:

- **Unit A:** Voltex® 120 V Plug-In Hybrid Electric HPWH.
- **Unit B:** Voltex® 240 V AL Smart Hybrid Electric HPWH.

Pros/Cons

- **120 V (Unit A)** *Pros:* standard outlet, compact/top connections, leak alerts. *Cons:* lower capacity and slower recovery (limited wattage).
- **240 V (Unit B)** *Pros:* higher UEF (3.88), faster recovery (higher wattage), advanced anode, iCOMM connectivity. *Cons:* requires 30 A dedicated circuit and professional installation.

TABLE I
COMPARISON OF 120 V AND 240 V A.O. SMITH HPWH UNITS

Item	Unit A: 120V HPWH	Unit B: 240V HPWH
Capacity	66 gallons	66 gallons
Type	Electric	Electric
First Hour Rating	76 gallons	82 gallons
Pipe Connection	Top only	Top and front
UEF	3.0	3.88
Annual Savings	∼ $118 vs. gas water heater	∼ $600 vs. conventional electric WH
Smart Connectivity	Bluetooth / Wi-Fi	A.O. Smith iCOMM app
Temperature Setpoint	Up to 140°F	Up to 140°F
Electrical Connection	60 Hz, 1-phase, 15 A shared	50/60 Hz, 1-phase, 30 A dedicated
Maximum Wattage	900 W	4500 W
Anode Type	Standard anode	Advanced anode (enhanced corrosion protection)

B. Operational Issues with the 240V HPWH

The 240 V unit was intended for side-by-side comparison with the 120 V model. Since installation, it has repeatedly failed to reach the 120 °F setpoint (remaining near

979-8-3315-8138-1

100–110 °F), despite the 45–55 min warm-up time specified by the manufacturer. Consequently, the 240 V unit was excluded from analysis, and the comparative study is deferred.

C. Operating Modes of the 120V HPWH

The 120 V unit supports *Heat Pump*, *Electric*, and *Vacation* modes, and the control panel indicates the active mode (Fig. 1). Outlet temperature is continuously logged. Real-time power is measured with an eGauge meter installed on the 120 V supply at the Syracuse Center of Excellence; measurements are streamed and accessible via a web interface for remote monitoring and archival.

Heat Pump: Highest efficiency; operates best at 37–120 °F ambient. If ambient is out of range or inlet $< 55°$F, the unit switches to Electric until conditions recover.

Electric: Uses resistance elements; recommended for high demand or unfavorable ambient conditions.

Vacation: Lowers tank temperature for extended inactivity. Default duration is 7 days.

Fig. 1. Control panel interface for the 120 V HPWH.

D. Related Work

Residential heat pump water heaters (HPWHs) have been studied from the perspectives of performance characterization, control strategy design, and demand-side flexibility. Sparn et al. [1] conducted a comprehensive laboratory evaluation of integrated HPWH units from multiple manufacturers, reporting efficiency, recovery time, and performance across wide ambient/inlet temperature ranges and producing detailed Coefficient of Performance (COP) maps that inform simulation and control design. Complementing this empirical lens, Hepbasli and Kalinci [2] reviewed heat pump water heating systems across technologies, working fluids, and configurations, highlighting how cycle design, compressor characteristics, and heat exchanger choices drive COP variation and practical deployment. Chua et al. [3] surveyed broader advances in heat pump systems, emphasizing improved cycle architectures and hybrid configurations that raise COP and extend applicability from industrial processes to residential water heating; the review also underscored installation constraints and climate sensitivity that matter for field performance.

On the evaluation and control side, Larson et al. [4] combined laboratory testing with energy-use estimation to show mode-dependent behavior and sizable savings potential, arguing for realistic draw profiles and representative environments in assessments. Carew et al. [5] analyzed load flexibility using the HPWHsim engine with California climate and CBECC-Res draw data, comparing timer-based to price-optimized strategies and validating against lab data; their results indicated that predictive, price-aware setpoint control can materially reduce peaks and costs. Together, these studies establish a robust foundation for HPWH performance modeling, COP-aware operation, and grid-interactive control, but most prior work centers on laboratory characterization, physics-based simulation, or rule-based scheduling.

The present study extends this line by introducing high-resolution field-style data from an A.O. Smith 120 V plug-in hybrid HPWH operating under dynamic usage profiles and grid-interactive conditions, and by benchmarking data-driven, short-horizon forecasting methods (tree ensembles, graph attention networks, Transformers, and few-shot/meta-learning). By fusing empirical telemetry with modern learning models, we target mode-aware prediction of power under varying thermal states and ambient conditions, enabling adaptive, data-driven scheduling for residential demand response and providing a reproducible dataset–methodology pair to support future HPWH control research.

II. Test Plan

Prior work reports that a three–to–four person U.S. household uses about 151–200 gal/person/day [6]. We adopt 180 gal/day as the design basis and distribute it across end-uses using room-level averages from [6]–[9]. The resulting profile emphasizes morning and late–afternoon peaks with light mid–day activity, reflecting common residential routines.

- Toilet: 18.5 gal/person/day (≈3 gal/flush)
- Washing machine: 15 gal/person/day (≈40 gal/run)
- Shower: 11.6 gal/person/day (≈7 gal/min)
- Faucet: 10.9 gal/person/day
- Dishwasher: 1 gal/person/day (≈20 gal/run)

Building access at the Syracuse Center of Excellence (CoE) is 09:00–17:00; draw events are mapped to this window to emulate a typical day while enabling on–site supervision (total ≈ 180 gal). Table II specifies the hour–by–hour allocation of activities and average volumes; micro–variations (e.g., ±10–20% duration jitter and randomized start offsets within each hour) are injected to avoid perfectly periodic patterns and to expose models to small timing perturbations.

This schedule serves as the baseline for scenarios with setpoint changes, mode switches, and hybrid transitions. We collect ten days of data—seven for training and three for evaluation—stratified by setpoint and operating mode to ensure

979-8-3315-8138-1

TABLE II
AVERAGE WATER USE MAPPED TO 09:00–17:00

Hour (24h)	Tasks	Avg. (gal)
09:00–10:00	Shower (10 min) + toilet (3 flushes)	60.5
10:00–11:00	Breakfast + shower (5 min) + light dishwashing	33.0
11:00–12:00	Handwashing + washing machine	12.0
12:00–13:00	No activity	0.0
13:00–14:00	Cooking + dishwashing + toilet (1 flush)	11.5
14:00–15:00	Minimal activity	2.0
15:00–16:00	Washing machine + general use	25.0
16:00–17:00	Faucet + dishwashing + full bathtub + toilet (2 flushes)	36.0

TABLE III
HPWH DATASET BY SETPOINT AND MODE: 7 TRAINING DAYS, 3 MIXED-MODE EVALUATION DAYS.

Set	Setpoint (°F)	Mode
Train #1	130	Heat Pump
Train #2	135	Heat Pump
Train #3	125	Heat Pump
Train #4	130	Electric
Train #5	135	Electric
Train #6	125	Electric
Train #7	50 (default)	Vacation
Eval #1	130	Electric (half day) + Heat Pump (half day)
Eval #2	120 (half) + 135 (half)	Heat Pump (half) + Electric (half)
Eval #3	130	Electric (half day) + Heat Pump (half day)

coverage of Heat Pump, Electric, and mixed days (Table III). Each day uses second–level telemetry (power; ambient, inlet, outlet temperatures; active mode) with synchronized timestamps from the eGauge meter and controller logs. Data quality control removes implausible spikes due to sensor dropouts; short gaps ($\leq$5 s) are linearly interpolated, and longer gaps are masked. Because several disturbances (e.g., setpoint steps) are discrete in time, we apply a light linear interior–point (LIP) smoothing to produce continuous trajectories for learning while preserving event timing.

The campaign yields $\sim$175 k training and $\sim$50 k test samples for the ML-based HPWH twin. For learning, we construct sliding windows (sequence length 60 s; 1 s stride for training, 5 s for evaluation to reduce overlap leakage). Standardization statistics are computed on the training set only and then applied to validation/test splits. To isolate generalization, we hold out entire days for testing rather than random pointwise splits; this preserves temporal dependence and prevents leakage across contiguous sequences. This protocol provides a reproducible basis for comparing algorithms under steady segments and abrupt transients while reflecting realistic field operation within the CoE access window.

III. EVALUATION RESULTS

We evaluate short-horizon power forecasting for the instrumented 120 V HPWH using second-level telemetry under realistic draw schedules. Data are segmented into fixed windows (sequence length 60 s) and split by day into train/validation/test as described earlier; all scalers are fit on the training portion only. Models are tuned on the validation set with early stopping and then assessed on the held-out test days. We report pointwise errors and summarize behavior during steady segments and abrupt transients to highlight robustness and data-coverage effects.

A. Evaluation Algorithms

We train and compare a total of eight forecasting models to build a digital twin of the HPWH and assess its response to disturbances (setpoint/mode changes, abrupt draws). All simulations were performed on a workstation equipped with an Intel Core i7-12700K CPU, 32 GB RAM, and an NVIDIA RTX 4070 GPU Models receive standardized inputs and predicts power (and other targets) at one-second resolution.

The model suite spans tree-based, meta-learning, graph-based, and sequence architectures: XGBoost (gradient-boosted trees); first-order MAML for few-shot learning (rapid adaptation from limited samples); a Graph Attention Network (GAT) that learns feature-wise dependencies via multi-head attention; a Feature-Tokenizer Transformer (FT-Transformer) that treats features as tokens; and four time-series baselines—LSTM, GRU, 1D-CNN, and a vanilla Transformer. Hyperparameters are chosen to balance accuracy and stability (typical settings: model width 64, 2–3 layers, 0.1–0.3 dropout, Adam optimizer with learning rate 10^{-3}, sequence length 60 for sequence models); early stopping prevents overfitting.

B. Evaluation Metrics

To evaluate model performance, we employ three commonly used regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Normalized RMSE. MAE measures the average magnitude of errors, providing an intuitive sense of prediction accuracy in the same units as the target (likewise for RMSE). However, NRMSE does not have a unit, meaning it is only a scaler out of 100. Formally, for each target j on a test set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_{\mathrm{te}}}$ with predictions $\{\hat{\mathbf{y}}_i\}$, the error metrics are defined as:

$$\mathrm{MAE}_j = \frac{1}{N_{\mathrm{te}}} \sum_{i=1}^{N_{\mathrm{te}}} \left| y_{i,j} - \hat{y}_{i,j} \right|, \tag{1}$$

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{N_{\mathrm{te}}} \sum_{i=1}^{N_{\mathrm{te}}} \left(y_{i,j} - \hat{y}_{i,j} \right)^2}, \tag{2}$$

$$\mathrm{NRMSE}_j = \frac{\mathrm{RMSE}_j}{\max_i y_{i,j} - \min_i y_{i,j}}. \tag{3}$$

When reporting a single overall figure, metrics are averaged across all m target variables; for example:

$$\overline{\mathrm{MAE}} = \frac{1}{m} \sum_{j=1}^{m} \mathrm{MAE}_j. \tag{4}$$

C. Results

The predictive performance of all models was evaluated on the HPWH dataset using the metrics described earlier. The results are summarized and discussed below.

Figures 2 and 3 show the predicted versus actual HPWH power for XGBoost, FSL, GAT, and FTT. Each model exhibits distinct characteristics. XGBoost trains quickly and performs well in steady-state regions but tends to smooth or lag during abrupt transitions. FSL improves adaptability to unseen patterns, though its outputs may show short-term fluctuations around sharp changes. GAT leverages attention to capture temporal dependencies, yielding smoother predictions but struggling with sudden load spikes or drops. The FTT model maintains long-term contextual consistency and mitigates drift, but its performance depends strongly on data quality, and it requires a higher computational cost.

All models were trained on a relatively narrow dataset, which naturally limits their accuracy. The modest performance observed here does not reflect a fundamental limitation of the algorithms themselves; accuracy is expected to improve as the dataset grows to cover broader operating conditions. Overall, these results highlight trade-offs between robustness, adaptability, efficiency, and data dependence, and suggest that richer datasets could unlock further potential.

Figure 5 provides a combined visualization of the prediction errors across the first four models, offering a clearer comparison of their ability to track the baseline trajectory. In addition, Figures 4a–4d illustrate the deep learning architectures (LSTM, GRU, 1D CNN, and Transformer), each displaying characteristic behavior when reproducing HPWH load dynamics.

The LSTM model benefits from memory cells that capture sequential dependencies and tracks long-term consumption trends well, but it oversmooths sharp transients and underestimates rapid switching events. GRU offers a more compact representation and reduced training complexity; however, it also struggles with abrupt spikes and may amplify prediction noise. The 1D CNN effectively detects localized patterns and produces sharp responses near transitions, though its limited receptive field leads to mismatches when long-term dependencies dominate. The Transformer, leveraging self-attention, captures global context and preserves long-range dependencies, yet is sensitive to rare extreme events and may exaggerate deviations during sudden changes.

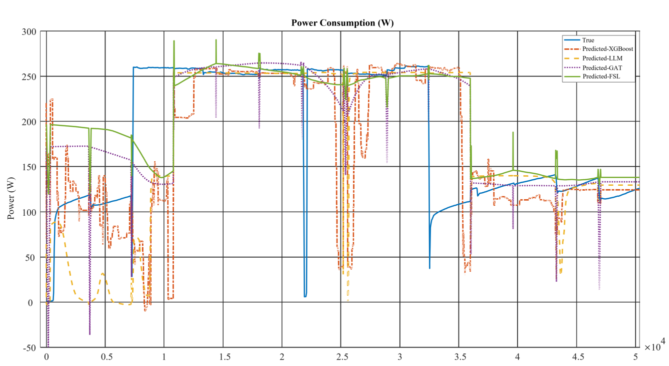


Fig. 5. Box plot of the number of positions sent per iteration using this scheme

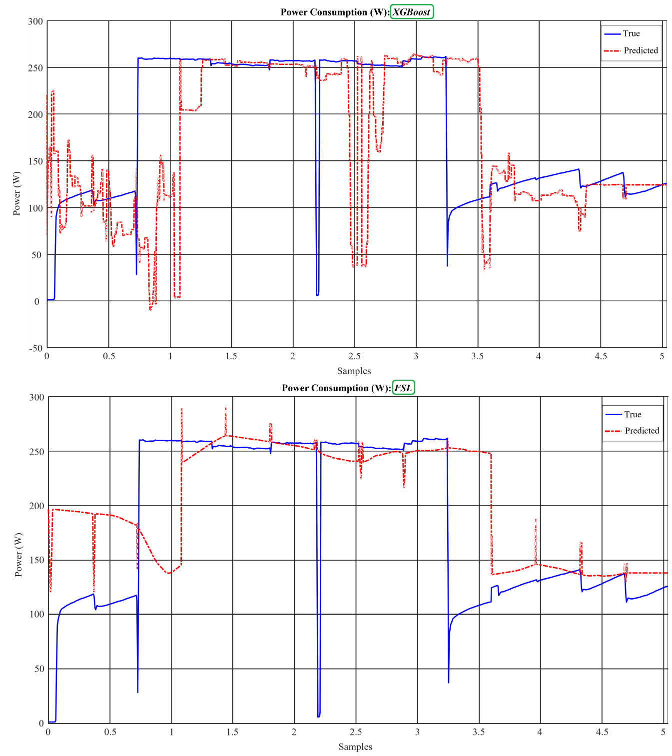


Fig. 2. HPWH power prediction: XGBoost (top) and FSL (bottom).

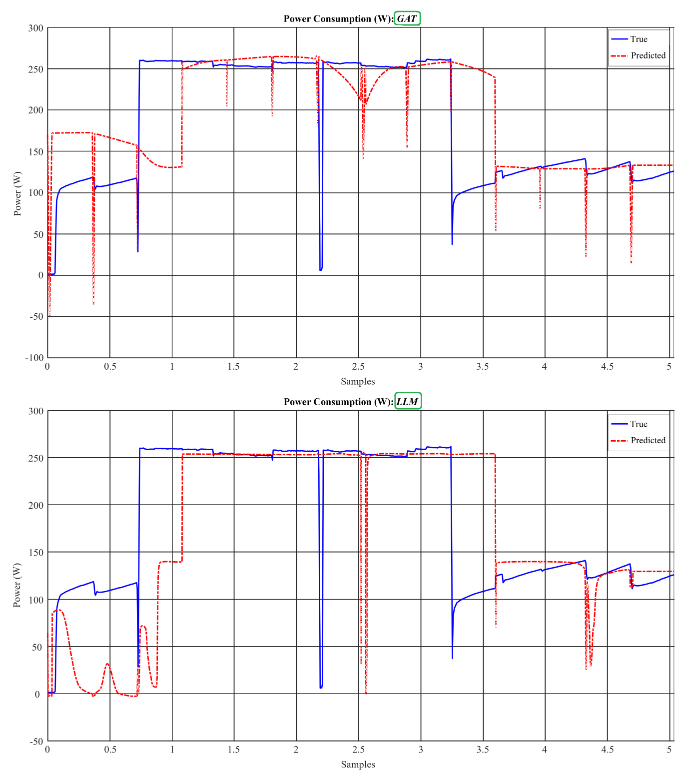
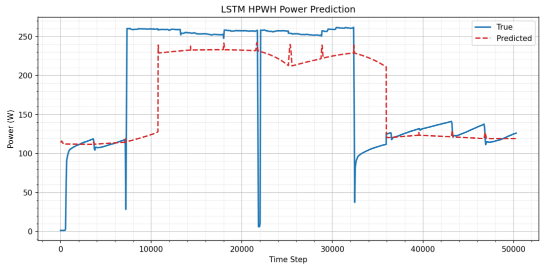


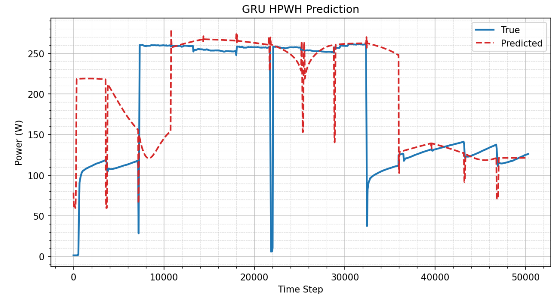
Fig. 3. HPWH power prediction: GAT (top) and FT-Transformer (bottom).

Overall, Table IV presents the evaluation results of the

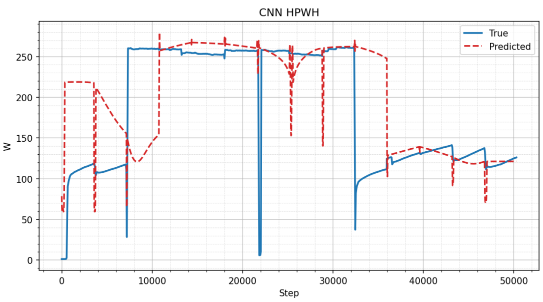
979-8-3315-8138-1



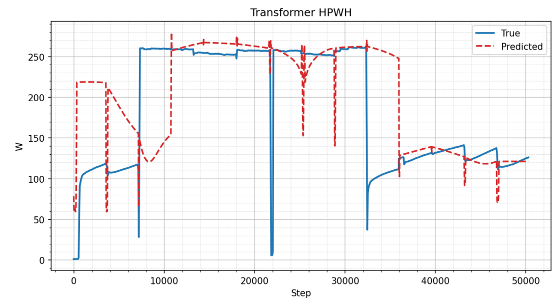
(a) LSTM



(b) GRU



(c) 1D CNN



(d) Vanilla Transformer

Fig. 4. Predicted vs. actual HPWH power for deep learning models: LSTM, GRU, 1D CNN, and vanilla Transformer.

models used on the test set. The results demonstrate that each algorithm exhibits specific advantages and limitations depending on the error metric considered. Among all, GAT and FSL achieved better results, reflecting their ability to handle diverse patterns despite limited data.

TABLE IV
EVALUATION RESULTS OF DIFFERENT MODELS ON THE TEST SET

Model	**MAE (W)**	**RMSE (W)**	**NRMSE (%)**
FSL	31.54	58.31	23
GAT	29.84	55.73	21
FTT (LLM)	33.21	65.92	27
XGBoost	35.18	70.29	29
LSTM (early stopped at epoch 11)	34.63	56.93	22
GRU (early stopped at epoch 12)	35.11	62.82	24
1D CNN (early stopped at epoch 12)	34.41	60.69	23
Vanilla Transformer (early stopped at epoch 15)	39.66	70.17	26

IV. CONCLUSION AND FUTURE WORK

This work presents an experimental, data-driven evaluation of a residential 120 V heat pump water heater (HPWH) under realistic draws using second-level power, temperature, and mode telemetry. We benchmarked tree ensembles, Transformers (including FT-Transformer and a vanilla encoder), graph attention networks, few-shot/meta-learning (MAML), and recurrent/conv nets (LSTM/GRU, 1D-CNN). Attention-based and meta-learning models delivered the best short-horizon errors, while tree and sequence baselines remained competitive in steady regimes; accuracy declined for abrupt transients, reflecting data sparsity and model bias. Two findings are central: (i) operating mode strongly drives energy use—electric-resistance–dominated days consume far more than heat-pump–dominated days—supporting mode-aware demand response; and (ii) simple persistence/moving-average baselines are necessary context for reported gains. Next steps include multi-season data, mode-conditioned metrics with uncertainty intervals, lightweight feature-importance analyses, and COP estimation from measured flow and temperature rise. Resolving the 240 V unit issues will enable side-by-side comparisons and hybrid rules-plus-prediction control for time-of-use load shifting.

REFERENCES

[1] B. Sparn, K. Hudon, and D. Christensen, "Laboratory performance evaluation of residential integrated heat pump water heaters," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2014.

[2] A. Hepbasli and Y. Kalinci, "A review of heat pump water heating systems," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 6-7, pp. 1211–1229, 2009.

[3] K. J. Chua, S. K. Chou, and W. Yang, "Advances in heat pump systems: A review," *Applied energy*, vol. 87, no. 12, pp. 3611–3624, 2010.

[4] B. Larson, M. E. Frankel, M. Logsdon, and D. Baylon, *Residential Heat Pump Water Heater Evaluation: Lab Testing & Energy Use Estimates*. Bonneville Power Administration, 2011.

[5] N. Carew, B. Larson, L. Piepmeier, and M. Logsdon, "Heat pump water heater electric load shifting: A modeling study," *Ecotope, Inc., Seattle*, 2018.

[6] A. Scavetta, "Average water usage in the united states," https://www.aquasana.com/info/average-water-usage-in-the-united-states-pd.html, 2025, [Online; accessed 2025-11-17].

[7] U.S. Geological Survey, "How much water do you use at home?" https://water.usgs.gov/edu/activity-percapita.php, 2025, [Online; accessed 2025-11-17].

[8] Erie County Water Authority, "Water use: Typical residential consumption," https://www.ecwa.org/my-water/water-use/, 2025, [Online; accessed 2025-11-17].

[9] South Florida Water Management District, "South florida environmental report," https://www.sfwmd.gov/, 2025, [Online; accessed 2025-11-17].

979-8-3315-8138-1