Energy Technology
Generation, Conversion, Storage, Distribution
www.entechnol.de

Characterization of Aggregated Building Heating, Ventilation, and Air Conditioning Load as a Flexibility Service Using Gray-Box Modeling

*Peter Kohlhepp, Lutz Gröll, and Veit Hagenmeyer**

Integrating large amounts of volatile renewable power into the electricity grid requires ancillary services (ASs) from multiple providers including flexible demand. These should be comparable by uniform and efficiently evaluable performance criteria. The objective is to characterize the technical flexibility of aggregated building heating, ventilation, and air conditioning (HVAC) under different operating conditions. New bounds of flexible power and holding durations, accordingly pay-back power and recovery times, and ramping rates are derived, using a new gray-box model of stochastically actuated aggregations of thermostatically controlled loads (TCLs) that can serve as well for load control. New closed formulas of the expected switching temperatures are derived using survival processes and hazard functions. This ex-ante characterization enables fast decision tools for AS feasibility testing and planning by demand aggregators, as it neither relies on simulation or optimization, nor on the identification and clustering of unit-level parameters. The estimates are explored in a sensitivity study of urban-level heat pump heating with respect to six key input factors. A case study using dynamic regulation signals from Pennsylvania–New Jersey–Maryland (PJM) demonstrates the benefit, in terms of tracking precision, of the refined energy measures over pure energy or power capacity bounds.

1. Introduction

A large part of the final energy demand in industrialized countries serves for the heating, ventilation, and air conditioning (HVAC) of buildings,[1] and the fraction provided electrically, e.g., through heat pumps, is increasing. These electric HVAC loads provide the grid with fast mechanisms to balance uncontrollable and unpredictable feed-in from intermittent renewable generation (i-RES), e.g., wind and photovoltaics (PV), as power consumption can be switched or modulated quickly (requiring no mechanical inertia), while the primary HVAC function is maintained using the thermal inertia in the background. Compared with generator plants with a steady fuel supply, the heat stored in buildings or tanks provides limited flexible energy. Therefore, the main target of flexible HVAC is short-term demand response (DR).

Working examples now exist in many parts of the world including Europe, the USA, and China.[2–6] Large commercial buildings have been providing dynamic regulation services on USA energy markets since 2011 and get payed for performance (Pennsylvania–New Jersey–Maryland [PJM][7]). Many small residential systems pooled together, especially thermostatically controlled loads (TCLs), can provide similarly good load balancing services.[8,9] However, they need an aggregation interface due to their low individual consumption, and aggregators' business models are still being challenged.[10] These doubts point to several larger problems, one being the imprecisely quantified value of residential DR for the grid compared with batteries (stationary or mobile) or to backup plants (fuel-based or storage-based), especially in future scenarios with ever higher shares of i-RES. On the other hand and beyond doubt, the flexibility costs in electricity systems are rising.[11]

Our study aim is to characterize the aggregated flexibility from end uses with thermal constraints, subsumed not only under "HVAC," such as domestic heating, separate water heating, air conditioning including ventilation, but also refrigeration and freezers. We refer to flexibility services (FSs)[12] as adjustments made to compensate for residual load due to forecasting errors or failures and compared with a previous state of knowledge implemented, for example, in a balanced schedule. The adjustment is to inject into, or absorb from the grid extra power, or to modify the consumption accordingly. The generic term includes ancillary services (ASs) defined in regulatory frameworks and initiated by the responsible grid operators to solve specific problems, e.g., to stabilize the grid frequency after an imbalance.

This article paper focuses on short-term FS/AS, especially real-time tracking and balancing of residual load such as

Dr. P. Kohlhepp, Prof. L. Gröll, Prof. V. Hagenmeyer
Institute for Automation and Applied Informatics
Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen 76344, Germany
E-mail: veit.hagenmeyer@kit.edu

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/ente.202100251.

DOI: 10.1002/ente.202100251

exemplified by the PJM dynamic regulation signals,[7] and also contingency support, e.g., taking specified load steps or ramps.

Criteria to characterize the flexibility technically include energy and power capacity, ramping speed, response times, response granularity and accuracy, recovery and payback needs,[13,14] tracking performance,[7] and reliability of services.[15] Trading and procuring FS/AS on the basis of abstract requirement specifications as proposed in the studies by Bondy[16] and the EU project SmartNet[17] requires multicriteria fact sheets to compare different FS supplier technologies, which seem to be lacking. At the operational level, DR aggregators need quick estimation and decision tools for feasibility testing, to match the specific flexibility (multicriteria) requirements with the capabilities under contract, before bidding into a flexibility market and before calculating detailed cost-optimal operating schedules.

2. Problem Definition, Related Work, and Contributions

We narrow down the scope and sharpen the requirements and assumptions of our study on DR characterization as follows. **Figure 1** provides an overview of the envisaged system architecture.

a) *Grid focus*: We seek technical criteria to characterize the flexibility limits. That is, the values or amounts should not rely on existing electricity markets and pricing mechanisms to allow for a technology-open comparison of different sources of flexibility, such as stationary or mobile batteries or flywheel storage compared with DR. The results should help shape future flexibility markets, and the criteria be relevant mainly for balancing responsible operators to assess how helpful the FS/AS from flexible demand are for transmission or distribution grids. Still, the criteria should be measured at the flexible resource itself, at the interface of requests and load responses.[18]

b) *Criteria depth*: We critically address the common criteria of flexibility performance, maximum power and energy deviation. Services such as secondary frequency response and short-term operating reserves impose requirements more stringent and specific than these, e.g., how long certain load deviations can be sustained.[19–22] As shown, holding durations do not trivially follow from a single energy bound. Certain tasks, such as voltage regulation in distribution grids, require analytical results on tracking errors under varying operating conditions.[23] In addition, to obtain a tailored flexible counterpart to intermittent solar and wind power, the dynamic ramping capability as well as the recovery and payback characteristics are crucial and need to be quantified.

c) *Load types, data knowledge*: We consider large pools of primarily residential buildings with their HVAC systems as the "units" (Figure 1, left). Unlike large commercial buildings that operate individually in markets and often have detailed and validated thermoelectrical models created in connection with a building energy management system (BEMS), the residential parameters at a community or city scale are often unknown or highly uncertain and strongly heterogeneous. At the unit level, only contractual information such as thermal constraints (limits) and power ratings may be reliable. During operation, the aggregate load sum is the only measurable response to the control signals issued to shape the load curves.

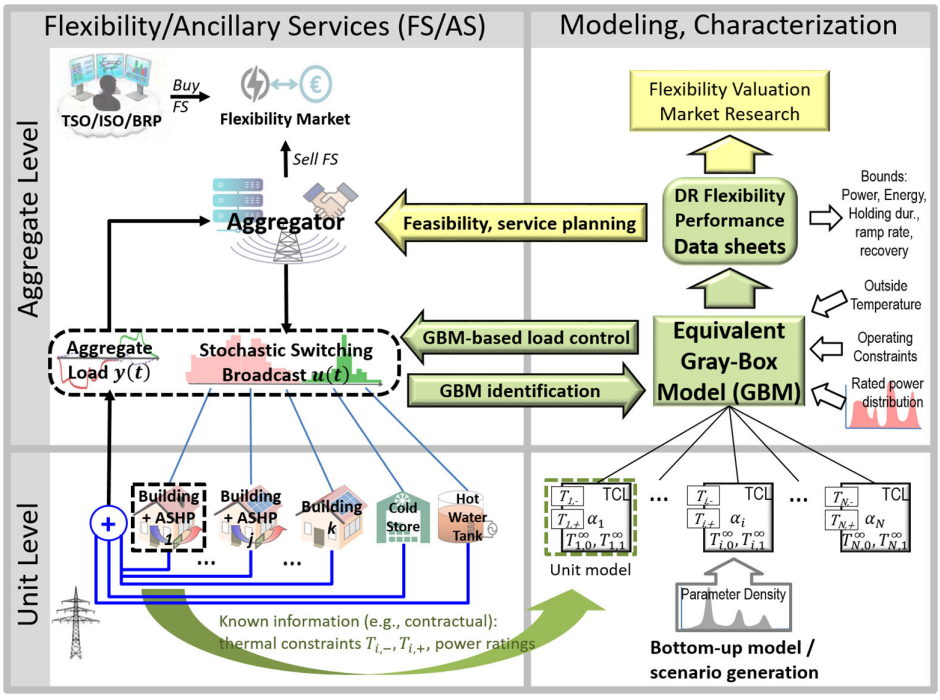

Figure 1. Aggregated flexible HVAC demand: system architecture and grid integration (left), and modeling tools (right). The green arc on the bottom symbolizes HVAC appliances being formally modeled as on/off controlled TCL units; i.e., the thermal end-use constraints are translated into switching points $T_{i,-}$, $T_{i,+}$; power ratings go into a separate distribution (on the right). Actual HVAC units might be more complex than TCL, for instance, admit continuous load modulation. The further TCL parameters α_i (drift rates), $T^{\infty}_{i,0}$, $T^{\infty}_{i,1}$ (limit temperatures) will be discussed in Section 3.1.

d) *Control interface*: When providing short-term FS/AS, we assume the units to be economically invisible, hidden by an aggregator (Figure 1, left). Unit-level consumption is not scheduled or economically dispatched but controlled in a closed loop and in real time by the aggregator. We therefore focus the achievable performance of aggregated services, not the unit properties. Assuming binary (on/off) controllable units, we analyze in particular randomized load control (stochastic switching, SSW) in addition to hard switching or programmable thermostats. The load is modified by specifying switching probabilities or rates or by selecting random subsets of units, possibly combined with random delays.[24–27] We want to characterize the flexibility achievable through stochastic actuation without assuming any specific control algorithm.

e) *Ex ante estimates*: Decision and planning tools for aggregators to check whether a service request is feasible with the contracted resources and to scale participation should provide performance limits quickly and easily, possibly those with reduced accuracy. Model-based estimates (of controlled load aggregations, in our case) can be categorized as "ex ante" or "ex post." The former estimates are obtained in closed form based on analysis of the model itself, the latter only by repeated simulation and/or optimization in the time domain with appropriately varied input signals and initial values, and subsequent extraction of performance data. Cost curves[28] and flexibility functions[29] that characterize the response to penalty signals are prominent examples of the ex post approach.

We study stochastic control (SSW in (d)) for three reasons. 1) SSW methods fully entrust to the local level, the units, to maintain their constraints; in particular, SSW satisfies all thermal end-use requirements by design. 2) SSW mitigates the oscillating dynamics that is due to synchronized hard switching, set point changes of thermostats, or energy refill of the population (payback/rebound). 3) SSW enables decentralized control schemes through broadcasting that require a minimal measurement and communication overhead[26,30,31] and meet requirement (c).

One possible practical barrier to mention is that thermostat controllers—be that domestic refrigerators or invertible heat pumps with a thermostat-guided compressor loop—respond to stochastic signals and make random choices. Though simple enough, these extensions lie beneath the optimization level of BEMS and are, to the best of our knowledge, not state-of-the art in most smart thermostat designs today.

2.1. Related Work

Flexibility characterization of DR as a research topic gained momentum with the proposal of energy storage and conversion models, such as the energy hub concept.[32] Thermal batteries (TBs)[33] are the probably best-known model to characterize electrical end-uses with quality constraints, especially TCL, and have been used for DR potential assessment, e.g., in California.[34] TBs characterize the admissible load profiles by specifying maximum upward and downward power deviations from the baseline consumption as well as maximum (thermal) energy deviation. The use of Minkowski summation lifts the TB description from units to aggregations. To express the overall flexibility again as a TB model, Hao et al.[33] derived upper and lower bounds, which diverge with the degree of population heterogeneity and assume the unit parameters to be known. This is not suitable for large load aggregations of incompletely modeled units (c).

Parallel to TB, convex polytopes and zonotopes have emerged as geometric flexibility characterizations in the power and energy dimensions.[35,36] The bounds defined by the union polytope are conservative to remain analytically tractable, and the geometric approach hardly scales with the problem complexity. Each new criterion or influencing parameter, such as the ambient temperature that shifts the baseline consumption as the reference point, creates a new dimension of analytical complexity, which conflicts with requirements (b) and (e).

Some efforts have been made in the TB framework to bound also the ramping speed of aggregate power and to consider lockout constraints that exclude TCL units temporarily from switching so to extend compressor life.[37–39] Bounds of the ramp rate[37] are derived from inventory equations of the TCL that are about to switch or have recently switched and require dynamic simulation to apply (e). The underlying analysis relies on the priority stack control (PSC), a centralized algorithm[33,37,39] that collects temperature measurements to determine a global switching priority, which becomes highest for the units closest to their regular switching points. The information architecture of PSC disagrees with our requirements (c,d). Ziras et al.[39] design the lockout periods so to minimize their adverse effects on the energy flexibility. The authors also compare the tracking performance of PSC with a stochastic controller in simulations. Other service types, e.g., contingency-type reserve actions that include the precharging of energy storage and the reconnection after service, have been analyzed in the study by Abiri-Jahromi and Bouffard.[40]

Coffman et al.[41,42] study lockout constraints as well but pursue a different goal named reference planning. A grid balancing authority successively extracts and assigns suitably shaped portions from a residual load signal to those flexible resources that can best handle these according to their capacities. This incremental matching problem is formulated as a convex optimization over a time horizon. Similarly as in Barth et al.,[43] the flexibility constraints describe TCL units mainly to optimally schedule them, less to characterize the overall performance that all flexible resources together can achieve on the overall residual load signal, contrary to (a,d,e).

While TB provide pure bounds on power or energy, the sustained power criterion (b) has been defined and used in the Belgian DR demonstration pilot LINEAR,[19,20,44] reasoning with standardized regulation profiles. Roossien[44] analyzed a special case of thermal buffers (deterministic control (d), time-linear heating/cooling, and homogeneous devices). More recently, Duong et al.[21] estimated expected durations under a coordination scheme that freezes TCL at specific temperatures and thus might cause short-cycling. Duration estimates have also been studied in the context of district heating systems (DHSs).[22] A higher-order ordinary differential equation (ODE) model with time delays describes the flow of hot water to the DHS-coupled buildings through pipes, a more complex system than considered in (c). Deriving flexibility boundaries therefore requires nonlinear optimization at each level of power deviation and does not admit a closed form (e).

By supplying capacity bounds as static constraints, TBs support planning and optimization tasks such as the economic dispatch of load and generation units.[35,41] To control and characterize load aggregations dynamically, TCL in fact lend themselves as a model and need no detour via TB or polytope descriptions, because the Fokker–Planck partial differential equation (FPE–PDE) raises their descriptive power to aggregations.[45–48] The FPE governs the evolution of temperature densities in time. Exploiting it for flexibility assessment, however, requires numerical PDE solutions (ex post analysis in the time domain (e)) and temperature state histograms as data representations (Markov chains[24,25,27]) (c). Several performance criteria have been assessed in this framework, e.g., granularity, ramping speed, or payback/recovery needs.[9,28,49–52] These results often characterize a specific load control algorithm developed and analyzed concomitantly (d), as discussed in the following text.

One important and indeed general characterization concerns the granularity[13] or quantization of load levels realizable with finitely many, the currently power-consuming TCL units. Assuming binary loads and a distribution of their power ratings, the aggregated load as a sum over random subsets is binomially distributed, and its standard deviation follows from the central limit theorem.[49,50,52,53] We exploit these results to factor out individual power ratings as a statistically independent random variable and rate load subsets as fractions of total rated power proportional to their size (normalized load factor).

Power and energy bounds are crucial in DR potential assessment studies,[34,54,55] some of which use TCL models for analysis. The study[55] of domestic power-to-heat potentials in Germany derives formulas for storage capacity, flexible power, and also holding durations (b), but fundamentally underestimates the latter by assuming that the units hold their desired states continuously and without interruption. Most estimates for DR potential assessment are too crude to serve as decision tools or data sheets for grid operators or aggregators (a).

Detailed and accurate performance estimates are often derived together with TCL control algorithm development (d). For example, Vrettos et al.[9] study the performance of refrigerator populations delivering primary frequency control (PFC) through stochastic switching and compare the impacts of various operating conditions, e.g., start-up load dynamics, thermostat resolution, and door openings on the load following accuracy. The evaluation is based on very detailed simulation studies ("ex post" (e)). Holding durations at specific power levels (b) are unknown, although the loads are challenged by signals with nonzero energy demand and might face energy depletion in several ways.

Ramping speed bounds (b) have not yet received very much attention. Hu et al.[51,56] derive bounds for TCL aggregations under centralized deterministic set point control. The authors rely on clustering algorithms for heterogeneous populations (c), on uniform temperature distributions, and on specific control assumptions (both the set points and the change rates of temperatures serve as control input). The bounds do not cover stochastic actuation (d). As they are asymmetric regarding upward and downward regulation and quadratic in the rated power, their theoretical and empirical justification is not obvious.

Tindemans et al.[26] develop a decentralized framework for SSW control, where each TCL unit independently targets its own reference load. Cooling and heating rates are controlled as intermediary variables. The authors derive envelope bounds for their algorithm "ex ante" (Equation 46 in the study by Tindemans et al.,[26] for cooling devices under load reduction), which are narrower than TB bounds of instant flexible power. The control framework addresses heterogeneous populations but uses solutions of the Fokker–Planck PDE which, however, governs homogeneous populations (b).[57] Trovato et al.[30] embed the concept into a leaky storage model and present an energy bound based on the mean population dynamics.[47] To what extent these bounds can be reached using explicitly stochastic or randomized control input (d) is not explicitly addressed. Related flexibility criteria of payback and recovery needs are discussed in the study by Trovato et al.[52] within a scheduling and unit commitment framework (b,d).

2.1.1. Summary of Knowledge Gaps

Roughly speaking, two avenues exist to characterize the flexibility of TCL load aggregates: TBs using geometric aggregation (TB[15,33–41]), and dynamic analysis of temperature densities or histograms based on the Fokker–Planck equations (FPE,[8,9,24–27,45–52,56]). While the TB approach captures only few criteria, the more versatile FPE requires a costly ex post analysis through simulation or optimization. The measurability or estimability of temperature densities appears to be a questionable assumption. Both ways rely on unit-level parameters that are difficult to identify, e.g., to cluster the units. The holding duration criterion (b)—in our view important, but yet to be demonstrated—and the power ramping dynamics have rarely been assessed and only for special cases. We also miss probabilistic performance (holding) guarantees for heterogeneous aggregates. Individual closed from bounds available from the TCL literature[26,51,52] depend on specific control algorithms and do not yet provide a coherent basis for creating performance data sheets. Especially, we miss a reachability analysis for SSW control in general.

2.2. Main Contributions

The main contributions of this article are summarized as follows. A) We derive novel formulas of the sustained flexible power for TCL aggregations under stochastic control and independently for direct load control. These imply further bounds of payback power, recovery duration, and maximum service frequency. Using survival processes, we analytically derive how the expected switching temperatures depend on the switching rates (control input). The change dynamics of switching temperatures corresponds to and provides a link to the reachable load deviations. B) By substituting the unit-level switching constraints (as hazard functions), we obtain a new coupled ODE of mean population temperature and aggregated load, denoted as a gray-box model (GBM, Figure 1, right). The three-state nonlinear GBM, independent of aggregate size, yields a new bound on power ramp rates (upward and downward) and suggests new algorithms for real-time stochastic load control that implement device-saving switching priorities locally. Using these controls, we can test how far the flexibility bounds are achievable. C) We establish distributions of the holding durations (HDDs) at any power level. The HDD order statistics then allows probabilistic guarantees (lower

bounds) of the flexible energy available at a given level of flexibility (power). The supply of flexible energy compared with the demand at high and low power levels decides how accurately a reference load can be tracked, as we demonstrate in a case study using the PJM dynamic regulation signals. Sufficient energy capacity (TB) alone does not guarantee accurate tracking, especially not in case of strongly heterogeneous populations with thermal diffusion (noise).

GBM in (B) are understood to be physics-based models of the aggregated load dynamics; in our case, of thermoelectric appliances with thermal constraints and derived from the FPE. Unlike white-box models, GBM need not be constructed bottom-up from detailed and validated building models. Compared with black-box (artificial neural network) models of electricity consumption,[58] GBM behavior is easier to understand and to diagnose. Due to its built-in specialization, the training and identification effort for useful models can be lower. Specifically, the initialization of the parameters in a GBM can make use of prior knowledge regarding building stock (for example, range estimates for thermophysical parameters) or user behavior.

Key influences to our GBM comprise ambient temperature, respectively, temperature-equivalent gains and losses, thermal comfort or safety intervals as the end-use quality constraints, control input in the form of stochastic switching rates, and heating/cooling rates that may (but need not) be specified through a parameter distribution. Furthermore, unit-level operating conditions such as the responsiveness to stochastic switching are specified by hazard functions (of the time in state).

The GBM coefficients for a fictitious aggregation can be generated by sampling from basic parameter distributions of the building and the HVAC stock (Figure 1, bottom right). By modifying these distributions, we empirically analyze the sensitivity of the characterization results to heat insulation, construction type, ambient temperature, and comfort constraints. To aggregate an existing population, on the other hand, a GBM with the appropriate coefficients will be identified directly from aggregated load traces under purposeful control excitation (nonlinear system identification), skipping the unit-level parameters.

As we make no direct use of the Fokker–Planck equations, i.e., require neither analytical nor numerical PDE solutions, we can characterize heterogeneous aggregates directly through transformations of probability densities (PDF) of input parameters. In particular, we require no prior similarity clustering of unit parameters. Second, obtaining crude ex ante flexibility bounds in closed form through Monte Carlo sampling is much faster than exploring-and-testing through simulation and optimization, which entails nested iteration loops over time and over skillfully varied initial conditions, controls, and disturbances. These are major expected benefits.

2.3. Outline of the Article

The remainder of this article proceeds as follows. Section 3 provides an overview of the methodical derivations and lays the foundations of the TCL theory used. Section 4 proposes the new aggregate model of electrical end-use processes with thermal constraints and presents new estimates of stationary or dynamic flexibility criteria, preferably in closed-form. The results will be empirically tested in Section 5 in several respects, verified in detailed simulation case studies, and applied to fast regulation service planning. Section 6 concludes our findings and suggests future work.

3. Overview and Preliminaries

In this section, an overview of the methodical derivations is provided and the foundations of the TCL theory used are lain in view of the following Section 4, where we derive new methods to obtain bounds for TCL aggregations sustaining a given load difference (flexibility) from the baseline load. Baseline is the load that assures the primary service quality against varying environmental demands. The flexible "energy-at-load-level"[59] forms the basis of further performance criteria.

Manageable and useful closed expressions or bounds, however, can be obtained mostly for special cases. Lacking a uniform methodology to analyze flexible energy-at-load-level, respectively, sustained or delayed flexibility, we derive several partial models (all probabilistic) that make different simplifying assumptions, starting with the simplest. To aid readers in understanding the model structure and logic, **Table 1** shows the main steps and results in Section 4.1–4.3. The table columns explain the TCL use cases and operating conditions, the flexibility properties and assertion types, the modeling assumptions and mathematical methods used. A short textual summary follows.

Section 4.1 describes individual load states (on/off) are controlled directly and the units selected assume and hold the target state for a certain period ("individually dispatchable," or ID units). We derive the expected flexible load sustainable for that time; the energy products are not constant. To obtain closed expressions, we need assumptions regarding the distributions of duty cycle durations and temperatures, which limit their practical use. We still use these results for later comparison with SSW populations.

Section 4.2 describes stochastic control transfers switching rates or probabilities to a set of TCL. Individual units are free to change states while collectively approximating the overall target. We model the unit behavior as survival processes with switching hazards that vary within duty cycles. As load differences are effected through changes (derivatives) of switching temperatures, respectively, control input, we obtain indeed TB-like energy bounds and derive the switching temperatures analytically. For this, we assume that control forces act instantly and neglect the population dynamics ("zero-order aggregation model" $AM^{(0)}$).

Section 4.3 describes that by substituting the hazard functions, a closed ODE system of mean population temperature and aggregated load dynamics results ("first-order aggregation model" $AM^{(1)}$), which bounds the ramp rates (power up/down) of SSW algorithms. The mean temperature equation connects load levels to holding durations in the form of a holding duration distribution (HDD). Its order statistics allow probabilistic guarantees conditioned on random TCL parameters. Power levels and durations yield variable products of flexible energy, which are in general smaller (tighter) than TB energy bounds and become even stricter with guarantees.

Table 1. Overview of characterization models and results in Section 4.

Section	Flexibility criteria	Operating conditions	Main results	Result type	Model assumptions	Mathematical methods
4.1 ID aggreg.	Flex. power cap. sustained	Coordination varying baseline $\bar{\gamma}^{\mathrm{bl}}(T_{\mathrm{amb}})$	Equation (10) Proposition 1	Expectation (time-varying)	Uninterrupted unit model UM$^{(0)}$ logn duration PDF uniform temp. PDF	Basic probability
4.2 SSW aggreg.	TB energy cap. recovery time service rate	$\bar{\gamma}^{\mathrm{bl}}(T_{\mathrm{amb}})$	Equation (17,21) Proposition 2	Upper bound (necess./ suff.)	SSW Aggr. model AM$^{(0)}$ steady state	Steady-state approx. (Equation (15))
Switching temperatures	$\bar{\gamma}^{\mathrm{bl}}(T_{\mathrm{amb}})$ SSW hazard, lockout	Equation (24,27) Proposition 3	Expectation	Aggr. model AM$^{(0)}$ SSW rate ftn.	Fokker–Planck PDE survival/hazard	Switching temperatures
4.3 SSW aggreg.	Energy at $\Delta\gamma$ holding times ramp rate	$\bar{\gamma}^{\mathrm{bl}}(T_{\mathrm{amb}})$	Equation (29,30,33,35)	Upper bound (necessary)	Aggr. model AM$^{(1)}$ SSW rate ftn.	Fokker–Planck PDE
Energy cap. Holding times	SSW hazard, lockout local control	Equation (37,38) Proposition 4	HDD (CDF/PDF) confidence levels	Aggr. model AM$^{(1)}$ continuous load steady-state	Order statistics	Energy cap. holding times

3.1. TCL Review

This section introduces the notation and collects facts and formulas about TCL populations needed in the following sections. We use the standard TCL model due to,[26,47] which consists of a bi-state (stochastic) heat ODE [60]

$$\begin{aligned} \mathrm{d}T(t) &= \alpha(T - T_s^{\infty})\mathrm{d}t + \sigma_{W,s} \cdot \mathrm{d}W(t),\ s = s(t) \in \{1, 0\} \\ T(t_0) &= T^{(0)} \end{aligned} \tag{1}$$

The process has one negative drifting rate $\alpha < 0$, two asymptotic temperatures $T_1^{\infty}, T_0^{\infty}$ approached in the limit $t \to \infty$, and a zero-mean diffusion term $\sigma_{W,s} \cdot dW(t)$ with variance $\sigma_{W,s}^2$. Units that heat up (respectively, cool down) in the active, power-consuming state are denoted active heating—AH (respectively, active cooling—AC) devices. State switching follows the thermostat hysteresis function, depending on the cases AH/AC

$$s(t) = \begin{cases} 1, & \text{if} \quad T(t) \begin{cases} > T_1 \ (\mathrm{AC}) \\ < T_1 \ (\mathrm{AH}) \end{cases} \\ 0, & \text{if} \quad T(t) \begin{cases} < T_0 \ (\mathrm{AC}) \\ > T_0 \ (\mathrm{AH}) \end{cases} \\ s(t^-), & \text{otherwise} \end{cases} \tag{2}$$

where T_s denote the temperatures at which the TCL switch into the according state $s = 1$ (on) or $s = 0$ (off); $T_1 > T_0$ holds for AC and $T_1 < T_0$ for AH devices. $s(t^-)$ is a shorthand notation for the left-side limit of the state function. T_0 and T_1 border the interval I of thermal safety of, for example, hot water prepared in a tank, or the thermal comfort of heated or cooled indoor air (quality-of-service). This interval $I = [T_-, T_+] := [\min\{T_0, T_1\}, \max\{T_0, T_1\}]$ of width $D := |T_1 - T_0|$ is defined by the end user (of each appliance) and constrains the thermal flexibility. Unlike most TCL literature, we require no specific set point to control but regard any point in I as acceptable.

Electric power uptake switches between zero and the maximum P^{TCL} (kW), close to the nameplate or rated power (idealized[9]) and is connected to the thermal power through a factor $\eta(t)$ (heat efficiency COP).[32] To simplify notation and calculations, we ignore the absolute power ratings of units and aggregations and consider normalized load factors in $[0, 1]$, i.e., we average several on and off cycles and devide by P^{TCL}. We use the standard TCL (1), (2) as an abstract unit model for any type of on–off-controllable HVAC device or subsystem with a thermal constraint.[61]

By virtue of (1), (2), TCL run alternating renewal processes[62] of on and off periods denoted duty cycles. The probability of a unit i being in state s_i in the long term is approximated by the ratio of durations $\overline{\tau}_{i,s}$ of on and off periods, which yields an approximate base load or baseline factor[54,62,63]

$$\begin{aligned} &\Pr\,(s_i = s) = \tfrac{\overline{\tau}_{i,s}}{\overline{\tau}_{i,s} + \overline{\tau}_{i,1-s}} \in [0, 1], \text{ where} \\ &\overline{\tau}_{i,s} = \tfrac{1}{\alpha_i} \ln \tfrac{\overline{T}_{i,1-s} - \overline{T}_{i,s}^{\infty}}{\overline{T}_{i,1} - \overline{T}_{i,s}^{\infty}}, \quad s \in \{0, 1\} \end{aligned} \tag{3}$$

The overline notation signifies stationary values ($t \to \infty$); $\overline{\tau}_{i,s}$ denote time-averaged durations spent by unit $i (1 \le i \le N)$ in state s. Temperatures may vary due to ambient conditions but stay bounded. It is also possible to average over units—separately for AH and AC to prevent sign cancellation—and approximate the base load factor, which is then time-varying, by interchanging quotient and averaging (this time, overlines signify population means, and $s(t)$ the state of a random unit)

$$\begin{aligned} &\bar{\gamma}^{\mathrm{bl}}(t) := \Pr\,(s(t) = 1) \approx \tfrac{\overline{\tau}_1(t)}{\overline{\tau}_0(t) + \overline{\tau}_1(t)}, \text{ where} \\ &\overline{\tau}_s(t) = \tfrac{1}{\alpha} \ln \tfrac{\overline{T}_{1-s}(t) - \overline{T}_s^{\infty}(t)}{\overline{T}_s(t) - \overline{T}_s^{\infty}(t)},\ s \in \{0, 1\} \end{aligned} \tag{3a}$$

The times calculated in (3), (3a) are valid only for limit temperatures strictly outside I so that the logarithm is negative but finite. Units i that are both upward and downward flexible (participating[33]) satisfy $T_{i,0}^{\infty} < T_{i,1} < T_{i,0} < T_{i,1}^{\infty}$ (AH) or $T_{i,1}^{\infty} < T_{i,0} < T_{i,1} < T_{i,0}^{\infty}$ (AC) and are collected in the set M^{pcp}. In a heterogeneous population, some units may never reach or leave their comfort bands. Permanently on or off units contribute with a base load of one, respectively, zero.

Following the notation in the study by Roossien,[44] we define the aggregated power flexibility by the symbols $\mathcal{F}^+(t,\tau)$, $\mathcal{F}^-(t,\tau)$ with activation time t and holding duration τ. For $\tau = 0$, the definition returns instant power capacity, the greatest possible load deviation from baseline. The symbol carries the sign of the residual load to cancel: $\mathcal{F}^+$ indicates load reduction to balance positive residual load and $\mathcal{F}^-$ load increase to balance negative residual load. However, $\mathcal{F}^\pm$ returns the difference $\Delta y := y - y^{\mathrm{bl}}$ at the consumer side, which assumes the opposite sign. A specific TCL unit i with binary load has flexibility 1, -1, or 0, depending on its state

$$\mathcal{F}_i^+(t,0) = \begin{cases} -1, & \text{if } s_i(t) = 1 \\ 0, & \text{else} \end{cases}; \mathcal{F}_i^-(t,0) = \begin{cases} 0, & \text{if } s_i(t) = 1 \\ 1, & \text{else} \end{cases} \tag{4}$$

Generally, the aggregated flexibility is queried in unknown states; the outcome $\mathcal{F}^\pm(t,\tau)$ is a random variable (RV). Instant flexibility comes from those participating units (in the set M^{pcp}) that are not in the target state in (3a)

$$\begin{aligned} E(\mathcal{F}^+(t,0)) &\approx -\tfrac{|M^{\mathrm{pcp}}|}{N} \cdot \tfrac{\overline{\tau}_1(t)}{\overline{\tau}_0(t)+\overline{\tau}_1(t)} \geq -\overline{y}^{\mathrm{bl}}(t) \\ E(\mathcal{F}^-(t,0)) &\approx \tfrac{|M^{\mathrm{pcp}}|}{N} \cdot \tfrac{\overline{\tau}_0(t)}{\overline{\tau}_0(t)+\overline{\tau}_1(t)} \leq 1 - \overline{y}^{\mathrm{bl}}(t) \end{aligned} \tag{5}$$

In the long term, due to the repetition of on and off periods, all phase positions appear roughly equally likely. Excluding an initial time interval of length τ_s^{lk} in state s from switching (lockout constraint) lets the available fraction drop to $(\overline{\tau}_s - \tau_s^{lk})/\overline{\tau}_s$ in the free steady state. If all units participate, the steady-state power capacity with lockout constraints becomes

$$\begin{aligned} E(\mathcal{F}^+(\cdot,0)) &\approx -\tfrac{\max\{0,\overline{\tau}_1-\tau_1^{lk}\}}{\overline{\tau}_0+\overline{\tau}_1} \geq -\overline{y}^{\mathrm{bl}} \\ E(\mathcal{F}^-(\cdot,0)) &\approx \tfrac{\max\{0,\overline{\tau}_0-\tau_0^{lk}\}}{\overline{\tau}_0+\overline{\tau}_1} \leq 1 - \overline{y}^{\mathrm{bl}} \end{aligned} \tag{6}$$

However, at some point in time t, it is still possible that all units are inside, respectively, all are outside their locked periods. Therefore, the instant power capacity bound (5) remains valid and can be achieved despite lockout. More accurate bounds available from the study by Sanandaji et al.[37] account for dynamic states but rely on simulation to infer the states.

TBs typically propose constant upper energy bounds (**Figure 2**); so the flexibility $\mathcal{F}^\pm(\cdot,\tau)$ tends to zero as $\tau \to \infty$. However, TCL populations may sustain small load deviations indefinitely and still meet their constraints by allowing a comfort interval $T_- \leq T \leq T_+$ and not imposing a strict set point. In contrast, we will show that TB energy bounds often overestimate the flexible energy available at large load deviations.

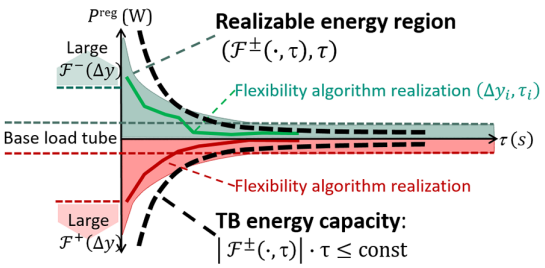

Figure 2. Energy capacity: load deviations and holding durations.

4. Aggregated Flexibility Model

4.1. Individually Dispatchable TCL

A query about the flexibility $\mathcal{F}(t,\tau)$ of ID units depends on two independent binary RV: being in the right state (1 for "+" and 0 for "−") for instant flexibility, and meeting the thermal constraints in the new state during the time interval $[t, t+\tau]$. Therefore, the expected sustained flexibility can be written as a product of two terms

$$E(\mathcal{F}^\pm(t,\tau)) = E(\mathcal{F}^\pm(t,0)) \cdot \Pr\left(T_{s,[t,t+\tau]} \subseteq I\right) \tag{7}$$

The second term decreases from one at $\tau = 0$ to zero at $\tau \geq \overline{\tau}_s$ and is roughly linearized in the study by Kohlhepp and Hagenmeyer[55] (Equation (9) therein). We take a different approach and capture the variations in duty cycle durations (RV ℓ) by a probability density function $f_{\ell,s}$. The expected flexibility follows by integrating over all duty cycle durations $c \geq \tau$

$$E(\mathcal{F}^\pm(t,\tau)) \approx E(\mathcal{F}^\pm(t,0)) \cdot \int_\tau^\infty \frac{c-\tau}{c} f_{\ell,s}(c)\,\mathrm{d}c \tag{8}$$

To see why approximation (8) is valid, consider that a unit will meet the thermal constraints in the new target state s for τ more time if and only if it switches no later than $c - \tau$, which applies to a fraction $(c-\tau)/c$ of all cycles of duration c, if cycle positions (phases) are uniformly distributed and heating/cooling linear in time ("zero-order" units). These units enjoy 100% flexibility, and the rest none. Integrating these quantities leads to the expected overall flexibility for duration τ. With increasing holding time τ, the integral, upper bounded by the tail of density $f_{\ell,s}$, quickly decreases to zero.

We further observe that load deviations could be sustained longer than for one duty cycle: not all units must hold their requested power state incessantly. Disjoint subgroups could be designated and allocated in a time-relayed fashion such that each one services a short interval $\delta \ll \tau$, as shown in **Figure 3**. Small δ then imply a higher availability due to (8) while the total power per subgroup shrinks to an average fraction $\approx \delta/\tau$. The product therefore has a maximum between the limiting cases of singleton groups ($\delta = \tau/N$) and the entire population ($\delta = \tau$).

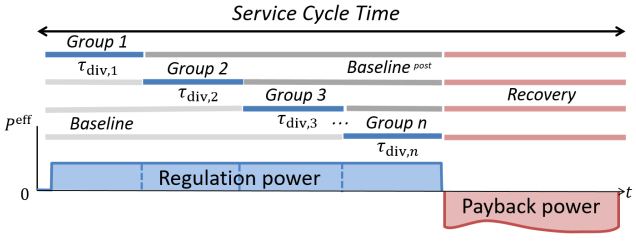

Figure 3. Relay scheduling diagram. Each TCL subgroup is engaged once in a service cycle, while the other groups consume baseline load (at a different temperature after servicing) denoted Baseline$^{\mathrm{post}}$. Recovery starts not until all groups have finished service.

The maximum flexibility depends on the duty cycle distribution. Empirical densities (histograms) collected from the first passage times of heterogeneous thermal processes are often well approximated by inverse Gaussian or by log-normal densities. We opt for the log-normal distribution that has been used before to analyze transient oscillations.[64] This leads to a new closed-form flexibility result.

Proposition 1: Assuming a population of zero-order ID TCL units with admissible and uniformly distributed temperatures for both load states at time t, the expected load difference sustainable for duration τ is

$$E(\mathscr{F}^{\pm}(t,\tau)) \approx E(\mathscr{F}^{\pm}(t,0)) \cdot \max_{0<\delta\le\tau} g(\tau,\delta), \text{ where}$$
$$g(\tau,\delta) := \frac{\delta}{\tau} \cdot \int_{\delta}^{\infty} \frac{c-\delta}{c} f_{\ell,s}(c)\,\mathrm{d}c \tag{9}$$

and $f_{\ell,s}$ denote the duty cycle duration pdf in states $s=0$ for "+" and $s=1$ for "−." If $f_{\ell,s}$ is log-normal with shape parameters μ and σ, the goal function g maximized in (9) for the internal service interval δ has a closed form

$$g(\tau,\delta) = \frac{\delta}{2\tau}\left(1+\operatorname{erf}\left(\frac{\ln(\delta)-\mu}{\sigma\sqrt{2}}\right)\right) - \frac{\delta^2}{2\tau}\cdot \mathrm{e}^{-\mu+\sigma^2/2}\cdot\left(1+\operatorname{erf}\left(\frac{\sigma^2+\ln(\delta)-\mu}{\sigma\sqrt{2}}\right)\right) \tag{10}$$

Proof: The first Equation (9) follows from (8), using the power reduction factor δ/τ for subgroups. The second Equation (10), goal function g in closed form, follows by substituting the log-normal density function (see Equation (11) below) into (9) and integrating by substitution of variables.[65]

The derivative $\partial g(\tau,\delta)/\partial\delta$ with appropriate limits for $\tau \to 0$ and $\sigma \to 0$ exists and can be zeroed uniquely to find the maximum. Other duration densities, e.g., empirical ones, are integrated numerically to maximize (9). An example calculation showing the impact of the density shape is shown in **Figure 4**.

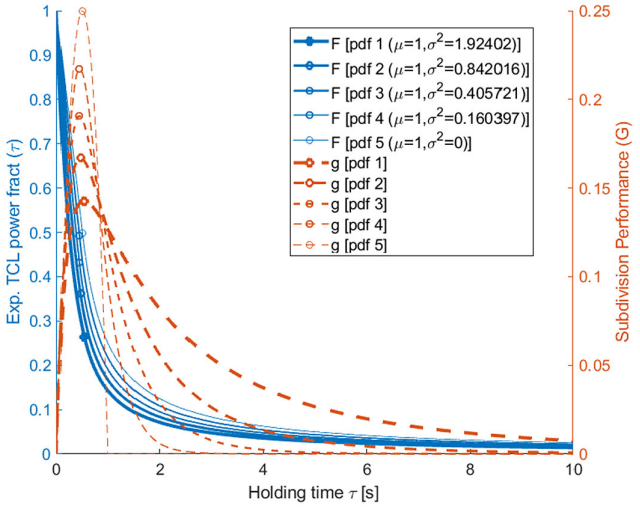

Figure 4. Flexible power $\mathscr{F}(t,\tau)$ (dark-blue curve) and subdivision performance g of an ID population for different duration PDFs. The optimum subgroup engagement δ is indicated by circles. For holding times $\tau > \delta$, $\mathscr{F}$ results from (9,10). Notably, power capacity slightly decreases with increasing variance.

The shape parameters μ and σ are related to expectation $\overline{\tau}$ and variance $\overline{\sigma}^2$ of the log-normal distribution as follows

$$f_{\ell}^{\log N}(\tau) = \begin{cases} \frac{1}{\sigma\tau\sqrt{2\pi}}\mathrm{e}^{-\frac{(\mu-\ln\tau)^2}{2\sigma^2}}, & \text{if } \tau > 0 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$
$$\overline{\tau} = \mathrm{E}(\tau) = \mathrm{e}^{\mu+\sigma^2/2},\ \overline{\sigma}^2 = \mathrm{Var}(\tau) = \mathrm{e}^{2\mu+\sigma^2}(\mathrm{e}^{\sigma^2}-1)$$

Some critical remarks are appropriate regarding the practical implications of this analysis of ID units:

Implementing the coordination scheme in Figure 3 correctly is not straightforward: the consumption of pausing TCL is to be controlled to stay at their baseline power, whereas storage temperatures differ but still obey the limits. Therefore, the estimated power capacity in (9) may not fully be realizable under ID control.

Approximately, uniform temperature distributions as required in (8) may prevail under baseline conditions, but no longer after service. Analyzing the time to recover the baseline conditions requires extra effort.[64,66]

Equation (11) proposes a log-normal distribution as a general model for the duration of duty cycles, whose parameters μ, σ depend on unknown TCL parameters and on the ambient temperature. Collecting measured durations is costly, and identifying the pdf shape from other measurements is not easy.

It will turn out that SSW populations require only average durations and no special analysis of rebound. Nevertheless, the ID performance results from (9) are valuable for comparison with SSW populations in Section 5.

4.2. Stochastically Controlled TCL

Stochastic switching causes TCL units to abort their duty cycles randomly and switch states before they reach the interval end points which act as a fallback position similar to a timed transition of a stochastic automaton (**Figure 5**). Stochastic switching shortens the duty cycles similarly as deliberate changes of the end points but in a randomized manner. The conditions that permit (guard) the stochastic transitions (yellow arrows in Figure 5) protect devices against too frequent forced switching (lockout constraints), but do not prevent thermostat switching. Thermal constraints are therefore obeyed as precisely as the

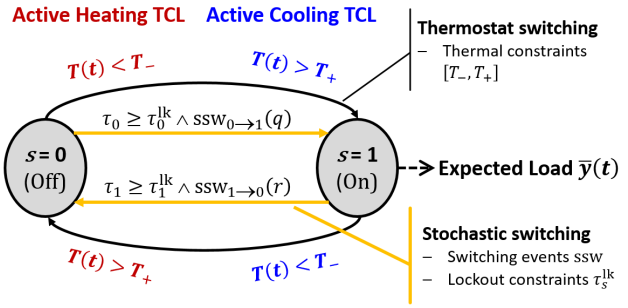

Figure 5. TCL automaton: τ_s is the time in state $s \in \{0,1\}$ and is reset by each transition into s, and τ_s^{lk} denote the lockout. The switching predicates SSW return random binary values: $\mathrm{SSW}_{0\to1}(q) \Leftrightarrow x \le 1-\mathrm{e}^{-qd\tau}$, $\mathrm{SSW}_{1\to0}(r) \Leftrightarrow x \le 1-\mathrm{e}^{-rd\tau}$, where $x \sim U_{[0,1]}$ is a uniform RV (Bernoulli trial) and $d\tau$ a discrete cycle time.

thermostat monitoring by the appliances permits; they are not violated due to stochastic switching.

4.2.1. Stationary and Zero-Order Bounds

For the first SSW bounds, we make use of steady-state temperatures and load factors, which exist due to bounded magnitudes and do not depend on initial values. Therefore, steady-state approximations apply to any single TCL as to the aggregation. In the steady state, the mean of all lower and all upper temperature switching points denoted $\overline{\mathrm{TSP}}_-, \overline{\mathrm{TSP}}_+$ approximates the mean population temperature $\overline{T}$ to second order; interested readers may refer to the Supporting Information for details. Similarly, substituting the TSP for T_s into the duration in (3) approximates the steady-state load factor $\bar{\gamma}$

$$\bar{\gamma} \approx \frac{\overline{\mathrm{TSP}}_1 - T_0^\infty}{\overline{\mathrm{TSP}}_1 - \overline{\mathrm{TSP}}_0 + T_1^\infty - T_0^\infty} \tag{12}$$

Equation (12) informs about the variations by moving the TSP positions inside the comfort interval I. The steady-state load is bounded from below and from above as follows, stated here for AH and similar for AC

$$\begin{matrix} T_0^\infty < T_- \le \overline{\mathrm{TSP}}_0 < \overline{\mathrm{TSP}}_1 \le T_+ < T_1^\infty \\ \underbrace{\frac{T_- - T_0^\infty + D_{\min}}{T_1^\infty - T_0^\infty + D}}_{[T_-,\ T_- + D_{\min}]} < \bar{\gamma} < \underbrace{\frac{T_+ - T_0^\infty}{T_1^\infty - T_0^\infty + D_{\min}}}_{[T_+ - D_{\min},\ T_+]} \end{matrix} \tag{13}$$

where the appliances require a minimum clearing distance $0 < D_{\min} \le \overline{\mathrm{TSP}}_1 - \overline{\mathrm{TSP}}_0 \le D$. The TSP intervals that obtain the minimal and maximal loads are shown under Equation (13). When the limit temperatures lie far outside a rather narrow comfort band, only a few percent of load difference are possible by shifting the TSP positions, far below the power capacity bound (5). These deviations can be sustained indefinitely without violating the thermal constraints and constitute a baseline interval (Figure 2). Bound (13) applies to heterogeneous aggregations using sample means, i.e., drawing values from distributions of $T_-, T_+, T_0^\infty, T_1^\infty$ and evaluating (13).

Example: Assume $T_0^\infty = 0$, $T_1^\infty = 50$, $I = [20, 22]$ (°C), $D_{\min} = 1$ K. With these values, the load factor varies between 0.404 and 0.431 (by only ≈3%); with a double-wide band $[19, 23]$(°C) between 0.370 and 0.451 (≈8%, at least). For comparison, raising the outside temperature by 3 K obtains $[0.346, 0.373]$, lowering it by 3 K: $[0.462, 0.490]$.

The question remains how to trigger the higher power deviations promised by capacity bounds (5), (6). Indeed, not a high switching rate itself causes peak load deviations, but its derivative or rate of change achieves it by skewing the TSP slopes and thereby distorting the on-to-off fractions, as shown in **Figure 6**. To see this, we make the following simplifying assumptions.

a) All upper and all lower TSPs are connected by two piecewise linear envelope curves $\mathrm{TSP}_+(t)$, $\mathrm{TSP}_-(t)$ which instantly follow the control trajectory (rate profile) and enclose a middle curve

$$\mathrm{TSP}(t) := (\mathrm{TSP}_+(t) + \mathrm{TSP}_-(t))/2 \tag{14}$$

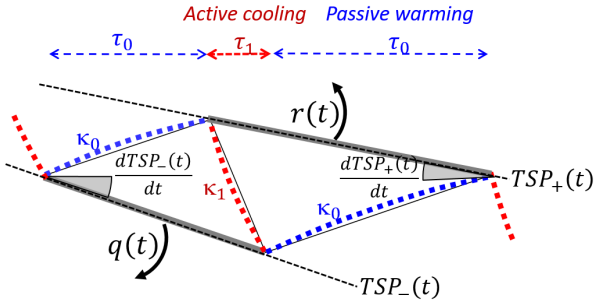

Figure 6. Derivatives of switching rates and switching temperatures control the aggregated load factor (illustration for AC).

that approximates the mean population temperature $\overline{T}(t)$ (zero-order aggregation model $\mathrm{AM}^{(0)}$).

$$\overline{\mathrm{TSP}}(t) \approx \overline{T}(t) \quad \text{as } t \to \infty\ (O(\epsilon^2)) \tag{15}$$

where $\epsilon := D/\min\{|T_1 - T_1^\infty|, |T_0 - T_0^\infty|\}$ relates the comfort band width to its distance from the limit temperatures that normally lie "far outside."

b) Unit models can be zero-order, abbreviated $\mathrm{UM}^{(0)}$, and have constant heating and cooling slopes $\kappa_s := (T_{1-s} - T_s)/\tau_s\ (s = 0, 1;\ \kappa_1 \cdot \kappa_0 < 0)$, or be first-order ($\mathrm{UM}^{(1)}$) as in (1), (2). The slopes are translated into first-order variables using the approximation

$$\kappa_s \approx \alpha\,(T_s - T_s^\infty),\ s \in \{0, 1\} \tag{16}$$

Using assumption (a) and geometric arguments shown in Figure 6, we approximate the combined impact of rising and falling TSP curves on the load factor using as "derivatives" the right derivative at the kink points.

Proposition 2: In a zero-order TCL aggregation model $AM^{(0)}$ the load difference (flexible power) is approximately proportional to the derivative of the mean temperature, if $|\dot{\overline{T}}(t)| < \min\{|\kappa_0|, |\kappa_1|\}$

$$\begin{aligned} \Delta\bar{\gamma}(t) &:= \bar{\gamma}(t) - \bar{\gamma}^{\mathrm{bl}}(t) \approx \frac{\dot{\overline{T}}(t)}{\kappa_1 - \kappa_0} \quad (\mathrm{UM}^{(0)}) \\ \Delta\bar{\gamma}(t) &\approx \frac{\dot{\overline{T}}(t)}{\alpha(T_0^\infty - T_1^\infty + T_1 - T_0)}. \quad (\mathrm{UM}^{(1)}) \end{aligned} \tag{17}$$

κ_1 and κ_0 have opposite signs. The leverage effect of temperature on the load is greater the slower the units heat up or cool down (unit: seconds per kelvin).

To sustain the load $|\Delta\bar{\gamma}|$ for duration τ, the mean temperature keeps changing with $\dot{\overline{T}}$ by (17). Integrating (17) obtains a temperature difference that must stay below D in magnitude. Due to lockout constraints, not the entire interval may be maneuverable. If the TSP must stay apart by a safety margin $\ge D_{\min}$ estimated, e.g., from the minimum duration τ^{lk} and/or the fraction ρ^{lk} being locked

$$D_{\min} \ge \max\{\max\{|\kappa_0|, |\kappa_1|\} \cdot \tau^{lk}, D \cdot \rho^{lk}\} \tag{18}$$

then (17) implies

$$|\Delta\overline{y}| \cdot \tau \le Q^{\mathrm{TCL}} :=\ \tfrac{D-D_{\min}}{|\kappa_1-\kappa_0|}\ (\mathrm{UM}^{(0)})$$
$$|\Delta\overline{y}| \cdot \tau \le Q^{\mathrm{TCL}} :=\ \tfrac{D-D_{\min}}{|\alpha|(|T_0^{\infty}-T_1^{\infty}|+D)}\ (\mathrm{UM}^{(1)}) \tag{19}$$

The (mean) temperature prior to service defines the normalized "thermal leeway" ϑ ($0 \le \vartheta_s(t_0) \le 1$, e.g., 0.5 for random activation times). The zero-order energy bound for SSW populations follows, using Q^{TCL} from (19)

$$\begin{aligned}&\tau_s^{\mathrm{FS}} \cdot |\Delta y^{\mathrm{FS}}| \le \vartheta_s(t_0) \cdot Q^{\mathrm{TCL}} \quad \text{Flexibility service (FS)}\\ &\tau_s^{\mathrm{PB}} \cdot |\Delta y^{\mathrm{PB}}| \le \vartheta_s(t_0) \cdot Q^{\mathrm{TCL}} \quad \text{Payback (PB)}\end{aligned} \tag{20}$$

The flexibility sign is immaterial in (20); the same product bound holds for flexible power Δy^{FS} and duration τ^{FS}, which moves the state of charge (SoC) in one direction as it holds for payback power Δy^{PB} and recovery duration τ^{PB} that goes in the opposite direction to restore the SoC. Realizability of the necessary upper bound under SSW control is discussed in Section 4.2.3.

Payback power that contributes to the area control error and burdens the grid can be reduced and accordingly more time can be allowed. The service frequency ω^{FS} at which a load deviation of Δy^{FS} can be sustained longest possible if the payback power in the subsequent recovery phase must not exceed $\Delta y^{\mathrm{PB}} \ll \Delta y^{\mathrm{FS}}$ is bounded as follows, due to (20)

$$\omega^{\mathrm{FS}}(\Delta y^{\mathrm{FS}}, \Delta y^{\mathrm{PB}}) \le \frac{|\Delta y^{\mathrm{FS}}| \cdot |\Delta y^{\mathrm{PB}}|}{|\Delta y^{\mathrm{FS}}| + |\Delta y^{\mathrm{PB}}|} \cdot \frac{|\kappa_1 - \kappa_0|}{D - D_{\min}} \tag{21}$$

Service frequency or service periods are further important performance criteria for FS.

4.2.2. Stochastic Control as Survival Process

Stochastic switching in this work follows Tindemans' semantics where the AS provider, e.g., an aggregator, and the participants share a continuous-time function which defines a switching rate at any time. In this form, switching rates can be included as bilinear control terms directly in the continuous Fokker–Planck–PDE,[26,28,47] the basic form of which describes the evolution of the temperature densities $f_1(T,t), f_0(T,t)$ of active, respectively, passive devices under heating and cooling:[45,46]

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t} f_s(T,t) = &-\underbrace{\tfrac{\partial}{\partial T}[\alpha(T-T_s^{\infty}) f_s(T,t)]}_{\text{Drift}} + \underbrace{\frac{\sigma_{W,s}^2}{2}\frac{\partial^2}{\partial T^2} f_s(T,t)}_{\text{Diffusion}}\\ &\underbrace{-r_{1-s}^{s}(T,t) f_s(T,t) + r_s^{1-s}(T,t) f_{1-s}(T,t)}_{\text{Stochastic switching}}\end{aligned} \tag{22a}$$

$$\begin{aligned}0 = &\underbrace{(T_s - T_1^{\infty}) f_1(T_s,t)}_{\text{Active flow out at } T_s} + \underbrace{(T_s - T_0^{\infty}) f_0(T_s,t)}_{\text{Passive flow in at } T_s}\\ &s \in \{0,1\}\end{aligned} \tag{22b}$$

Switching rates $r_{1-s}^{s}(r_1^0 \mathrel{\widehat{=}} q, r_0^1 \mathrel{\widehat{=}} r)$ normally affect TCL equally on all temperature levels. The boundary condition of conservation (22b) models the exchange between $f_1(\cdot,t)$ and $f_0(\cdot,t)$ due to thermostat switching at T_s.

In this form, stochastic switching can be regarded as a failure/survival process that is basically described by 1) A positive hazard function $h(t)$ causing components to fail at a certain rate; 2) A probability density function (pdf) $f^h(t)$ of failure times due to hazard h, denoted loss function;[67] 3) The related cumulative lifetime function (cdf) $F^h(t) = \Pr(\text{lifetime} \le t)$.

The link between survival and stochastic switching rates in the FPE (22) is discussed in Supporting Information. A constant switching rate in the FPE (22) indeed defines a valid constant hazard function. By specifying the switching hazard and evaluating the loss pdf, the switching temperature distribution or its moments can be determined.

Each switching event has a random exit temperature with unknown density $f_s^{\mathrm{tsp}}(T)$, linked to the exit time through a strictly monotonous transformation, the duration $\tau_s(T)$ to reach temperature T. The correspondence is $f_s^{\mathrm{sw}}(t) = f_s^{\mathrm{tsp}}(T(t))$ and $f_s^{\mathrm{sw}}(\tau_s(T)) = f_s^{\mathrm{tsp}}(T)$. Survivors of SSW switch thermostatically after time τ_s at temperature $T(\tau_s) = T_{1-s}$. The expected exit temperature out of state s depends on the entrance temperature θ, another random variable, e.g., due to the opposite switching, and reads using the definition of conditional expectation

$$\begin{aligned}E(\mathrm{TSP}_s|\theta) = &\int_{\theta}^{T_{1-s}} T \cdot \underbrace{f_s^{\mathrm{sw}}(\tau_s(T))\frac{\mathrm{d}\tau_s(T)}{\mathrm{d}T}}_{\text{Loss PDF of temperature}} \mathrm{d}T +\\ &T_{1-s}\underbrace{(1 - F_s^{\mathrm{sw}}(\tau_s(T_{1-s})))}_{\text{Survivors}},\ s \in \{0,1\}\end{aligned} \tag{23}$$

Equation (23) reflects the transition timing in Figure 5. To evaluate the TSP, the hazard function is specified as a unit-level switching policy which can be designed purposefully. For instance, the hazard remains zero or very low during an initial period, see **Figure 7**. In the further analysis, we consider two special cases of hazard functions and make appropriate simplifying assumptions of the unit models to obtain manageable closed expressions.

4.2.3. Constant and Linear Switching Hazard

The FPE with SSW control terms in (22) is tantamount to the use of a constant hazard function h. We consider the unit model $\mathrm{UM}^{(1)}$ with logarithmic mapping of temperatures onto durations but ignore the diffusion term in Section 4.2.2. Substituting into (23) and integration returns the conditional mean TSP

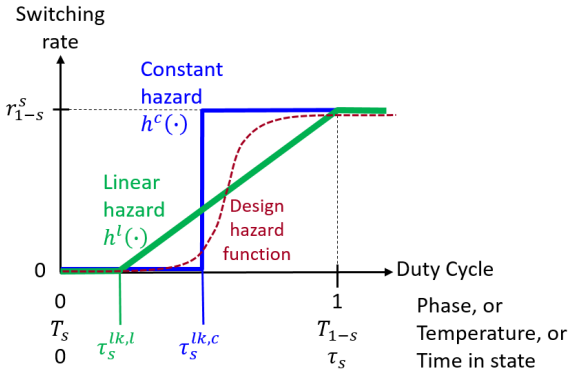

Figure 7. Switching hazard functions at unit level with initial lockout.

for the rate r^s_{1-s} (superscript "C" for constant hazard)

$$E\left(\mathrm{TSP}^{\mathrm{C}}_s|\theta\right) = T^{\infty}_s + \left(\frac{T_{1-s}-T^{\infty}_s}{\mathrm{e}^{\alpha\tau^{lk}_s}(\theta-T^{\infty}_s)}\right)^{-u_s} \frac{T^{\infty}_s-T_{1-s}}{u_s-1} + \frac{u_s \mathrm{e}^{\alpha\tau^{lk}_s}(\theta-T^{\infty}_s)}{u_s-1},\ s\in\{0,1\} \tag{24}$$

where $u_s := r^s_{1-s}/\alpha \le 0$ is the control signal, the SSW rate related to the thermal drift, and τ^{lk}_s $(0 \le \tau^{lk}_s \le \bar{\tau}_s)$ an initial period in state s blocked from switching.

If the entire cycle remains locked, the expressions $\mathrm{e}^{\alpha\tau^{lk}_s}(\theta - T^{\infty}_s)$ in (24) collapse into $T_{1-s} - T^{\infty}_s$ and result in the thermostat switching point $\mathrm{E}\left(\mathrm{TSP}^{\mathrm{C}}_s|\theta\right) = T_{1-s}$ for any control u_s and entry temperature θ; as well if SSW is absent ($u_s = 0$). For the opposite case of high SSW rates, in the limit, switching is expected at the temperature reached right after expiry of the period τ^{lk}_s

$$\lim_{u_s\to-\infty} \mathrm{E}\left(\mathrm{TSP}^{\mathrm{C}}_s|\theta\right) = T^{\infty}_s - \mathrm{e}^{\alpha\tau^{lk}_s}(\theta - T^{\infty}_s) \tag{25}$$

An alternative is a hazard growing within the duty cycle (Figure 7): devices should switch more likely as they approach their regular thermostat points and has been implemented in the priority stack control (PSC) algorithm.[33] The information how to compute switching probabilities from rates is available at the unit level through local measurements. A central algorithm collecting and sorting device temperatures is not obligatory; we outline a decentralized variant used for testing in this work in Appendix B. The linear hazard function reads

$$h(\tau) = \begin{cases} 0, & \text{if } 0 \le \tau < \tau^{lk}_s \\ r^s_{1-s}\frac{\tau-\tau^{lk}_s}{\bar{\tau}_s-\tau^{lk}_s}, & \text{if } \tau^{lk}_s \le \tau \le \bar{\tau}_s \end{cases} \tag{26}$$

To simplify expressions and obtain a hazard function that is linear also in device temperature, we assume $\mathrm{UM}^{(0)}$. Substituting (26) into the pdf $f^{\mathrm{sw}}_{\tau}(t)$ and integrating (23), the expected TSP can be calculated

$$\mathrm{E}\left(\mathrm{TSP}^{L}_s|\theta\right) = \theta^{lk} + \mathrm{sgn}\left(D^{lk}_{\theta}\right)\sqrt{\frac{\pi D^{lk}_{\theta}}{2u_s}}\cdot\mathrm{erf}\left(\sqrt{\frac{u_s D^{lk}_{\theta}}{2}}\right) \tag{27}$$

where $u_s := r^s_{1-s}/\kappa_s$ relate the SSW rate to the thermal slope in state s (1 or 0), and θ^{lk} denotes the point between entry temperature θ and T_{1-s} from where SSW takes effect; $D^{lk}_{\theta} := T_{1-s} - \theta^{lk}$ is the distance from T_{1-s} and is linear in the initial blocked fraction $\rho := \tau^{lk}_s/\bar{\tau}_s$ $(0 \le \rho \le 1)$ of duty cycles.

Equation (24) and (27) explains the effects of one-way switching, which is the most important case in practice, but are not yet applicable to bidirectional switching, because expectations and conditions mutually depend. The following general result assures the existence and uniqueness of the expected TSP pair as the limit of a contracting sequence.

Proposition 3: For any TCL unit that meets the assumptions of (24) for constant hazard, or (27) for linear hazard and $UM^{(0)}$, and for each pair of SSW rates $q \ge 0, r \ge 0$, there exist unique temperature values $\overline{TSP}_1, \overline{TSP}_0 \in [T_-, T_+]$ such that 1) $\overline{\mathrm{TSP}}_s = \mathrm{E}\left(\mathrm{TSP}_s \,|\, \overline{\mathrm{TSP}}_{1-s}\right)$, $s \in \{0,1\}$; 2) Either $T_1 < \overline{\mathrm{TSP}}_0 < \overline{\mathrm{TSP}}_1 \le T_0$ (AH) or $T_0 \le \overline{TSP}_1 < \overline{TSP}_0 < T_1$ (AC); 3) $\overline{\mathrm{TSP}}_s, \overline{\mathrm{TSP}}_{1-s}$ are the (unconditionally) expected switching temperatures for (q, r) in the steady state.

Numerical examples illustrating how the two antagonistic rates r and q shape the switching temperature surfaces $\overline{\mathrm{TSP}}_1(r,q)$ and $\overline{\mathrm{TSP}}_0(r,q)$ are provided as surface plots and discussed in the Supporting Information. Equation (24) and (27), and Proposition 3 provide several insights and benefits.

They lead to a constructive proof that the bounds (4) and (17) are reachable through (one-way) SSW. Indeed, the functions (24) and (27) connect the SSW rates with the expected TSP, while Equation (14), (15), and (17) provides the link to the load factor. By requesting some load deviation $\Delta\bar{y}$ within the bounds (4), the required change rate of mean population, respectively, mean switching temperature $\overline{T}(t), \overline{\mathrm{TSP}}(t)$ then yields a scalar nonlinear ODE for the rate profile $u_s(t)$ (starting with $u_s(t_0) = 0$) to achieve that load level.[68]

$$\Delta\bar{y}(t) \overset{(17)}{\approx} \frac{\mathrm{d}\overline{T}(t)/\mathrm{d}t}{\kappa_1-\kappa_0} \overset{\text{chain rule}}{\overset{(14)(24)(27)}{\approx}} \frac{\mathrm{dE}\left(\mathrm{TSP}_{\pm}(u_s)\right)/\mathrm{d}\,u_s}{2(\kappa_1-\kappa_0)}\cdot\dot{u}_s(t) \tag{28}$$

where $\mathrm{dE}\left(\mathrm{TSP}_{\pm}(u_s)\right)/\mathrm{d}\,u_s$ is found by differentiating the appropriate Equation (24) or (27) with respect to the appropriate control input $u_s(t)$ (rate $r(t)$ or $q(t)$).

Equation (24) and (27) helps to quickly assess (analytically, in closed form) the possible degradation of thermal comfort: we see at a glance how far a comfort band chosen initially wide is deformed due to SSW control effect.

They help assessing the temporary depletion of the TCL set available for load control as the thermal windows shrink to width $\overline{\mathrm{TSP}}_+ - \overline{\mathrm{TSP}}_- - D_{\min}$ compared with $\overline{T}_+ - \overline{T}_-$ initially; which reduces the power ramping ability (acceleration). This insight aids in developing simplified dynamic aggregation models.

4.3. New Gray-Box Model

4.3.1. Mean-State ODE System

A sudden step input of switching rate (**Figure 8**) finds units in all phases of their current duty cycle but affects only their remainders as predicted in the previous section. We now derive essential but minimal model extensions to capture such delays in the form of a coupled ODE system (first-order aggregation model $\mathrm{AM}^{(1)}$, or GBM). ODE of temperature mean and variance in homogeneous TCL populations have been derived by integrating the Fokker–Planck PDE over the temperature range.[47] We also model the aggregated load dynamics in the form of bilinear or trilinear terms that connect state with control variables, by substituting the switching hazard functions from Section 4.2.3. The mean-state ODE for a SSW population reads

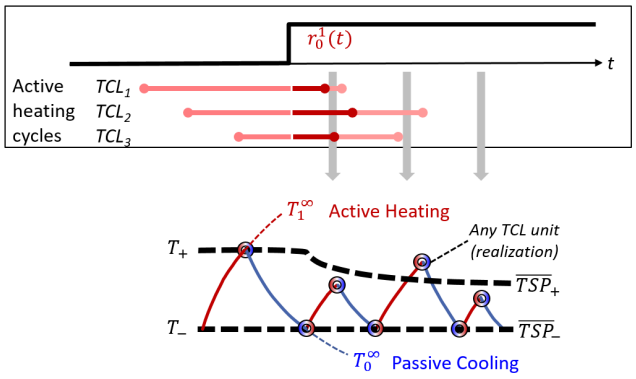

Figure 8. Illustration of SSW control in AM$^{(1)}$: A step rate (top) hits all active units, shortens their remaining duty cycles, and produces a delayed mean TSP$^+$ curve (thick dashed line).

$$\begin{pmatrix}\dot{\overline{T}}\\ \dot{\overline{y}}\end{pmatrix} = A\begin{pmatrix}\overline{T}\\ \overline{y}\end{pmatrix} + B_0(t) + B_u^{\{(C)|(L)\}}(\overline{T},\overline{y})\cdot u(t)$$
$$\text{where } A = \begin{bmatrix}\overline{\alpha} & \overline{\beta}_1 - \overline{\beta}_0\\ 0 & 0\end{bmatrix},\ B_0(t) = \begin{pmatrix}\overline{\beta}_0 + \chi_2(t)\\ \xi_{\{1|2\}}(t)\end{pmatrix} \tag{29}$$
$$B_u^{(C)} = (1-\overline{\rho}^{lk})\begin{bmatrix}0 & 0\\ 1-\overline{y} & -\overline{y}\end{bmatrix}$$

$$B_u^{(L)} \approx (1-\overline{\rho}^{lk})\begin{bmatrix}0 & 0\\ \overline{\vartheta}_1(1-\overline{y}) & -(1-\overline{\vartheta}_1)\overline{y}\end{bmatrix} \tag{30}$$

denote the system and control matrices for constant (C) and for linear (L) switching hazard, respectively; $B_u^{(C)}$ and $B_u^{(L)}$ have (products of) state variables as "coefficients." The control input $u(t) := (q(t), r(t))^T$ contains the switching rates on and off. We briefly explain the individual GBM elements; interested readers are referred to the Supporting Information discussing the connection to the governing FPE. We can write the coefficients of system matrices A and B_0 formally as device parameter means

$$\overline{\alpha} := \frac{1}{N}\sum_{i=1}^{N}\alpha_i,\ \overline{\beta}_s := -\frac{1}{N}\sum_{i=1}^{N}\alpha_i T_{s,i}^\infty,\ s\in\{0,1\} \tag{31}$$

but should identify them independently from load measurements if drift rates and limit temperatures are unknown. The first-row equation of mean temperature is—visually not immediately obvious—the first-order counterpart of (17)[69] and thus informs about the thermal energy budget of the population.

The disturbance terms $\xi_{\{1|2\}}(t)$ and $\chi_2(t)$ in matrix B_0 (29) remain as "leftovers" from the FPE reduction to a mean-state ODE and take the following forms

$$\begin{aligned}
\xi_1(t) &= [\alpha(T-T_1^\infty)\cdot f_1(T,t)]_{T=T_-}^{T=T_+} && \text{Pure drift } (\dot{\overline{y}})\\
\xi_2(t) &= \xi_1(t) + \tfrac{\sigma_{W,1}^2}{2}[\partial f_1(T,t)/\partial T]_{T=T_-}^{T=T_+} && \text{Diffusion } (\dot{\overline{y}})\\
\chi_2(t) &= \sum_{s=0,1}\tfrac{\sigma_{W,s}^2}{2}[T\,\partial f_s(T,t)/\partial T]_{T=T_-}^{T=T_+} && \text{Diffusion } (\dot{\overline{T}})\\
&\quad - \sum_{s=0,1}\tfrac{\sigma_{W,s}^2}{2}[f_s(T,t)]_{T=T_-}^{T=T_+}
\end{aligned} \tag{32}$$

where f_1, f_0 denote the time-varying temperature densities in the FPE (22). The term ξ_1 reflects the flow difference between active devices entering and leaving f_1 at the points T_+, T_- and may be actively suppressed through bidirectional switching, by adding small offsets $\epsilon_0^1, \epsilon_1^0$ to both rates r_0^1, r_1^0 only one of which is essential for each flexibility direction. We neglect the offset terms ξ_2 and $\chi_2(t)$ for diffusion in this work, lacking useful bounds of their magnitude, but remember them as possible causes of model mismatch.

The common factor $\overline{\rho}^{lk}$ in B_u in (30) has the physical meaning of a mean blocked fraction of all duty. Due to B_u, the system (29) is bilinear for constant and trilinear for linear hazard, in two state variables ($\overline{y}$ and the thermal leeway $\overline{\vartheta}_s$ remaining for a population to adopt state $s=1$ or 0), and in the control variable u; see Appendix Appendix B (B.2).

We emphasize the aggregated load dynamics $\dot{\overline{y}}$ in the second row of (30) as a new bound on ramping speed

$$\begin{aligned}
\dot{\overline{y}}^{(C)} &\approx (1-\overline{\rho}^{lk})\,[(1-\overline{y})\cdot q - \overline{y}\cdot r]\\
\dot{\overline{y}}^{(L)} &\approx (1-\overline{\rho}^{lk})\,[\overline{\vartheta}_1\cdot(1-\overline{y})\cdot q - (1-\overline{\vartheta}_1)\cdot\overline{y}\cdot r]
\end{aligned} \tag{33}$$

All quantities in (33) except q and r are in $[0,1]$.

4.3.2. Holding Time Distributions

Assuming that any initial $\overline{T}_0 \in \overline{I}$ can be reached and sustained ($\dot{\overline{T}} = 0$) by applying suitable baseline load $0 < \overline{y}^{bl} < 1$, the mean population temperature dynamics under the load difference $\Delta\overline{y}(t) = \overline{y}(t) - \overline{y}^{bl} \in [-\overline{y}^{bl}, 1-\overline{y}^{bl}]$ follows (29) (first line)

$$\begin{aligned}
\dot{\overline{T}} &= \overline{\alpha}\,\overline{T} + (\overline{\beta}_1 - \overline{\beta}_0)\,\Delta\overline{y},\quad \overline{T}(t_0) := \overline{T}_0 \in \overline{I}\\
0 &= \overline{\alpha}\,\overline{T}_0 + (\overline{\beta}_1 - \overline{\beta}_0)\,\overline{y}^{bl} + \overline{\beta}_0 + \chi_2
\end{aligned} \tag{34}$$

Therefore, the difference $\Delta\overline{y}$ can be held without violating thermal constraints at most for a duration τ^h which depends on the initial mean state $\overline{T}_0$ within a "mean" band $\overline{I}$, and on the population parameters $\overline{\alpha}, \overline{\beta}_{\{1|0\}}$

$$\tau^h \le \begin{cases}\frac{1}{\overline{\alpha}}\ln\left(\frac{D(\Delta\overline{y}) + \overline{T}_- - \overline{T}_0}{D(\Delta\overline{y})}\right), & \text{if } D(\Delta\overline{y}) > \overline{T}_0 - \overline{T}_- > 0\\ \frac{1}{\overline{\alpha}}\ln\left(\frac{D(\Delta\overline{y}) + \overline{T}_+ - \overline{T}_0}{D(\Delta\overline{y})}\right), & \text{if } D(\Delta\overline{y}) < \overline{T}_0 - \overline{T}_+ < 0\\ \infty, & \text{else}\end{cases}$$
$$\text{where}\quad D(\Delta\overline{y}) := (\overline{\beta}_1 - \overline{\beta}_0)\,\Delta\overline{y}/\overline{\alpha} \tag{35}$$

In the first case, T_- and in the second case T_+ will be violated. If all units are flexible (participate), any feasible positive or negative $\Delta\overline{y}$ can be achieved and sustained for a—often very short—period; see **Figure 9** for a geometric illustration. The flexible energy exploitable on a specific level $Q(\Delta\overline{y}) = \Delta\overline{y}\cdot\tau^h$ stays below the zero-order energy bound (20), because the positive/negative logarithm is strictly concave/convex, respectively. The energy amount decreases with the load difference $\Delta\overline{y}$.

The bound (35) holds regardless of how $\Delta\overline{y}$ is controlled. Moreover, the ODE (34) for the mean population dynamics equally describes any specific unit or subset if we impose on

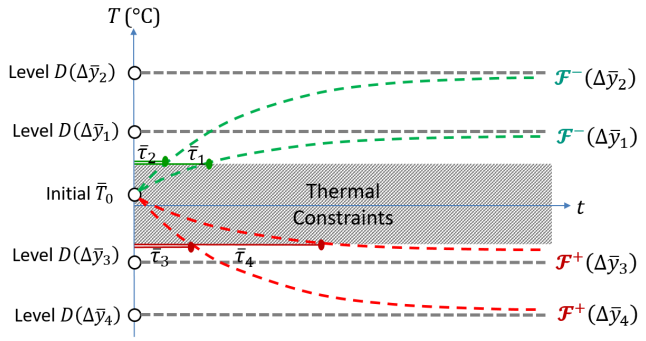

Figure 9. Holding durations for different load levels $\Delta\bar{y}$. The $\overline{T}$ curves all start at $\overline{T}_0$ inside a mean band $\bar{I}$ assumed reachable.

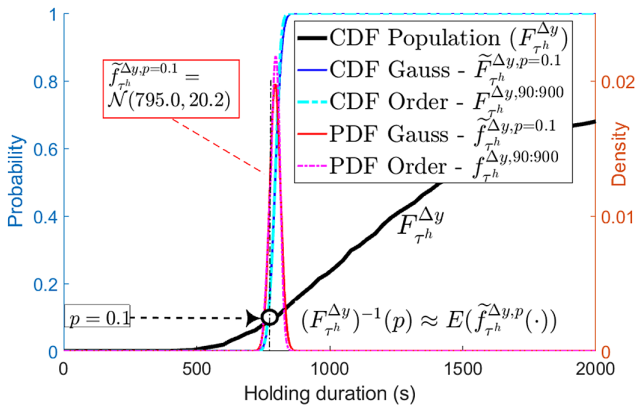

Figure 10. Empirical HDD example; see Example 1 in text.

it a real-valued (fractional) load level in the long term but abstract individual binary load states and durations away. This will indeed be the main application. The relation (35) transforms input parameter distributions into holding duration distributions. Using order statistics,[70] worst-case holding abilities can also be characterized with statistical guarantees. With the following proposition, we can determine the likelihood that a certain fraction of units will exceed a given duration.

Proposition 4: Assume a population of participating TCL with input densities $f_\alpha, f_\beta, f_{T_0}$ and initial temperatures $T_i \sim f_{T_0}$ that are all achievable with loads γ_i^{bl}, $0 < \gamma_{\min}^{\mathrm{bl}} \leq \gamma_i^{\mathrm{bl}} \leq \gamma_{\max}^{\mathrm{bl}} < 1\,(1 \leq i \leq N)$. Then the holding durations τ^{h} for $\Delta\bar{\gamma} \in (-\gamma_{\min}^{\mathrm{bl}}, 1 - \gamma_{\max}^{\mathrm{bl}})$ have a cumulative distribution defined by transformation (35)

$$F_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma}}(\tau) = \Pr(\tau^{\mathrm{h}}(\Delta\bar{\gamma}, \alpha, \beta, T_0) \leq \tau \,|\, \alpha \sim f_\alpha, \beta \sim f_\beta, T_0 \sim f_{T_0}) \tag{36}$$

If the quantile value $\xi = (F_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma}})^{-1}(p)$ for $p \in (0,1)$ exists and $F_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma}}$ is differentiable at ξ with positive derivative $f_{\tau^{\mathrm{h}}}(\xi) > 0$, then the holding duration undercut by $p \cdot N$ units at worst is an asymptotically normally distributed RV for large numbers N

$$f_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma},pN:N} \xrightarrow{d} \widetilde{F}_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma},p} = \mathcal{N}\left(\xi, \frac{p(p-1)}{N f_{\tau^{\mathrm{h}}}^2(\xi)}\right). \tag{37}$$

Example 1: Figure 10 displays the empirical HDD of a small heterogeneous population ($N = 900$) as the parent of the order statistics $F_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma},90:900}$, the holding time CDF undercut by at most 10% ($p = 0.1$) of the population. The order statistics is computed numerically using large (inaccurate) binomial coefficients and compared to its Gaussian approximation from Proposition 4.

The likelihood that $\Delta\bar{\gamma}$ can be sustained for τ or longer by at least $(1-p)N$ units is $1 - \widetilde{F}(\tau)$ with the Gaussian CDF $\widetilde{F}$ for $\widetilde{f}$. We denote fraction p as compliance level,[71] and define the flexible energy retrievable at load $\Delta\bar{\gamma}$ and compliance p using the Gaussian approximation

$$\begin{aligned} &\widetilde{Q}(\Delta\bar{\gamma}, p) := |\Delta\bar{\gamma}| \cdot \widetilde{\tau}(\Delta\bar{\gamma}, p) \\ &\text{where} \quad \widetilde{\tau}(\Delta\bar{\gamma}, p) := \mathrm{E}\,(\widetilde{F}_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma},p}(\cdot)) = \xi(p) \end{aligned} \tag{38}$$

The realistic (load-level-dependent) energy demand to track a regulation signal is straightforward to estimate (in Supporting Information) using the above $\widetilde{Q}(\Delta\bar{\gamma}, p)$ divided by the basic Q^{TCL} as an energy intensity factor.

A short numerical recipe summarizes how to practically apply Proposition 4:

Input	Load difference $\Delta\bar{\gamma}$
	Size $N > 1$; Compliance level $p\,(0 < p < 1)$
	TCL parameters $\alpha_i, \beta_i, T_i, T_{i,-}, T_{i,+}$
Output	Mean and variance of density $\widetilde{F}_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma}}$
Algorithm	1) Acquire the input densities $f_\alpha, f_\beta, f_{T_0}$.
	2) Build an empirical HDD (histogram $h_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma}}$, CDF $H_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma}}$): Sample input parameters using $f_\alpha, f_\beta, f_{T_0}$ and evaluate τ^{h} (35) (use $M \gg N$ samples).
	3) Estimate the mean $\xi = \min\{\tau : H_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma}}(\tau) \geq p\}$. Fit a density f, e.g., log-normal, to $h_{\tau^{\mathrm{h}}}^{\Delta\bar{\gamma}}$ and assure $f(\xi) > 0$. Repeat from Step 2 with more samples if necessary.
	4) Evaluate the Variance of $\widetilde{f}$ Using (37).

The critical step in practice will be providing suitable proposal distributions for the input parameters in Step 1. Regarding f_{T_0}, participants should keep the aggregator informed of their up-to-date tolerance bounds. Initial temperatures may then be assumed independent and uniformly distributed: $f_{T_0} \equiv \prod_i U_{[T_{-,i}, T_{+,i}]}$, lacking more specific information. For the remaining thermophysical parameters α and β, we see two options.

1) Forward (generating f_α and f_β for a fictitious building population—scenario analysis): sample from key parameters that characterize the building and HVAC stock and for which distributions may be easier to obtain, and calculate α_i, β_i from these. This will be discussed in Appendix A.

2) Inverse (identifying f_α and f_β of an existing population—service planning): estimate $\overline{\alpha}$ and $\overline{\beta}$ as coefficients of the GBM system matrix A in (29) from measured aggregated load under SSW control. Distribution shapes found in (1) or through data analysis may be adopted as templates for f_α, f_β and be rescaled to the identified means. Identifying also the variances from load profiles might be possible, however, a complete PDF shape rather not.

Example 2: Figure 11 highlights the key influences on the HDD, using the synthetic population[39] ($N = 10^5$ TCL in 10 random states) and a load step up $\Delta\overline{y} = 0.3$. We vary the population homogeneity by degree h: $h = 1$ means fully homogeneous; for $h = 0.9$ the parameters are varied ±10% about the reference values as in the study by Ziras et al.[39] The thick black line ($h = 0.7$) serves as our reference. We see from Figure 11 that the holding durations exceeded by 90% of units in the most homogeneous ($h = 0.9$, solid blue) more than double those in the most heterogeneous case ($h = 0.5$, solid orange). Ranking the input factors, different initial temperature concentrations in $[T_-, T_+]$ account for most, as shown by the dashed-dotted HDD curves. The HDD can also be shaped to a modest extent if the units target individual Δy_i adjusted to their holding abilities and achieve the flexibility $\Delta\overline{y} = 0.3$ in the mean (dark-purple dotted line).

5. Numerical Results

In Section 5.1 and 5.2, we explore the criteria and bounds from Section 4 and analyze their sensitivity to key input factors. The criteria are evaluated numerically in closed form and without simulation. Basically, we automatically generate TCL surrogate models from more basic parameters (distributions) that describe the built environment, HVAC equipment, and quality constraints, and then characterize their load flexibility properties through density transformation. Details of model generation are given in Appendix A. The further Section 5.3–5.5 address the realizability and accuracy of bounds, using independent methods for comparison, mostly detailed simulation. The comparison with work in the study by Ziras et al.[39] (Section 5.4) and the PJM case study (Section 5.5) serve to justify our refined energy criteria.

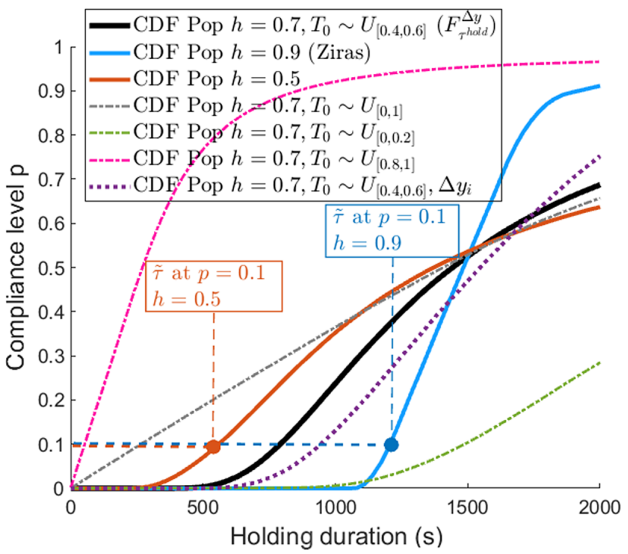

Figure 11. Different HDDs (CDF) for a target load $\Delta\overline{y} = 0.3$; see Example 2 in text.

5.1. Heat Pump Case Study

By aggregating whole-building models at a regional or city scale, we show how to obtain DR performance figures (in particular, holding durations of flexible power) very fast and ex ante. These capture various HVAC operating conditions including ambient temperature. Another goal is to compare the two independent estimation methods derived first for ID loads and later for SSW populations.

5.1.1. Data Description

The aggregation model is generated using the parameters in Table A1; further parameters are shown in **Table 2**. This TCL population serves as the reference for sensitivity experiments in Section 5.2. Acceptable temperatures assumed uniformly distributed lead to an initial SoC of 50%. Regarding tolerances, the actual comfort band is $I = [T_1, T_1 \pm D]$°C ("+" in this case of heating, see Section 3.1). Comfort bands in Table 2 are random parameters and assume building users that tolerate fairly wide temperature ranges of 4–6 °C. In the remainder, this reference configuration of heterogeneous building models is used in two variants, H-ASHP ("ASHP in heating mode") and C-ASHP ("ASHP in cooling mode"), each with individual thermostat set points, comfort tolerances, and outside temperatures. Even more control over the composition of populations by appliance classes and thermal constraints is possible through another configuration denoted M-TCL ("mixed-devices"). Five classes, i.e., space, heating and space cooling, separate water heating, domestic fridges, and freezer cabinets coexist in one population.[55] Each class

Table 2. Parameters of the reference configuration.

Meaning/Name	RV	Value[a)]	Unit
TCL type	N	AH (ASHP)	–
Switch on temperature T_1	Y	$U_{[17,21]}$	°C
Comfort band width D	Y	$U_{[4,6]}$	K
Initial SoC	N	0.5	–
Thermal diffusion σ_W	N	0.02	–
Lockout duration τ^{lk}	N	90	s
Specific heat capacity $\overline{C}$	Y	$\mathcal{N}^T(0.9, 0.2)$	kJ kg^{-1} K^{-1}
Heat transfer $\overline{U}$[a)]	Y	$\mathscr{H}_{[0.1,1.3]}$	W m^{-2} K^{-1}
Air change rate r_{ach}	Y	$\mathscr{H}_{[0.2,1.0]}$	h^{-1}
Design heat intensity $\overline{P}^{dli}$	Y	$\mathscr{H}_{[0.4,0.9]}$	W m^{-2}

[a)] $\mathscr{H}_{[0.1,1.3]}$ for heat transfer coefficients $\overline{U}$: Normalized counts of 12 bins of width 0.1 over the histogram domain [0.1, 1.3] were chosen (0.011, 0.032, 0.105, 0.211, 0.263, 0.105, 0.021, 0.021, 0.105, 0.011, 0.011, 0.105). $\mathcal{N}^T(\cdot,\cdot)$: shorthand for truncated normal distribution, conditional on range $\mathbb{R}_{>0}$.

independently provides positive and negative flexibility. A table of all parameter values is found in the Supporting Information.

5.1.2. Experiment Description

The available flexible power fraction is visualized as contour plots in a grid of ambient temperatures and holding durations in **Figure 12**. Direct (ID—top row) and stochastic actuation (SSW—bottom row) are compared. Using (10) for ID loads, the expected flexibility $\mathcal{F}^{\pm}(\cdot, \tau)$ is calculated for given holding durations.

The corresponding SSW values are computed as the minima obtained from the energy bound (20) and the holding time relation (35). Equation (18),(19) account for cycle availability (lockout), whereas (35) captures the decrease of retrievable energy with the load difference.

Regarding the ID versus SSW actuation type, **Figure 13** offers a compact visual comparison of flexible power in a bounded area of up to 40% higher, respectively, lower heat capacity values around the reference. In the left and middle diagrams, the sustainable power, averaging over durations $\bar{\tau}^{\text{hold}} \leq 400s$, is

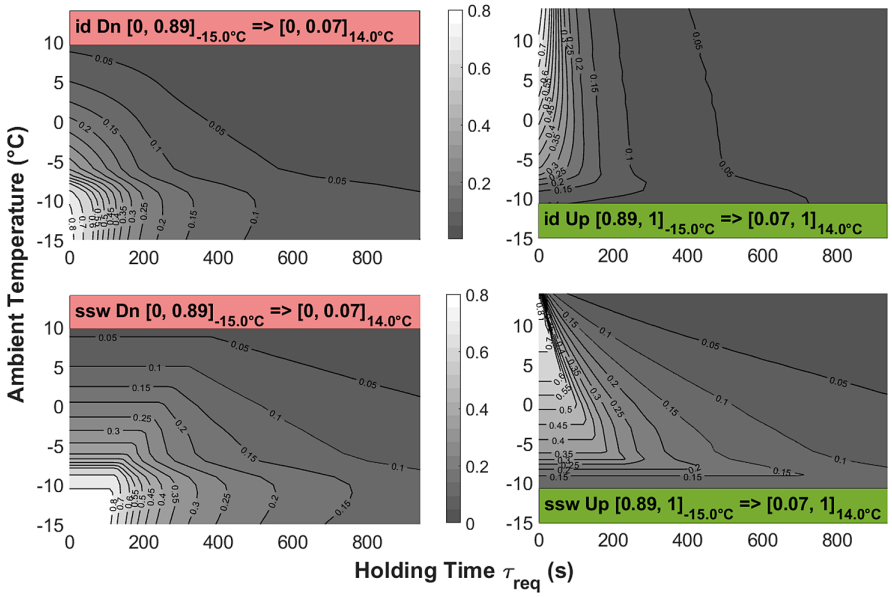

Figure 12. Flexible power capacity (normalized) of ASHP building population displayed as contour plots over a grid of holding durations and outside temperatures. Left column: downward, right: upward flexibility. Top row: ID, bottom: SSW.

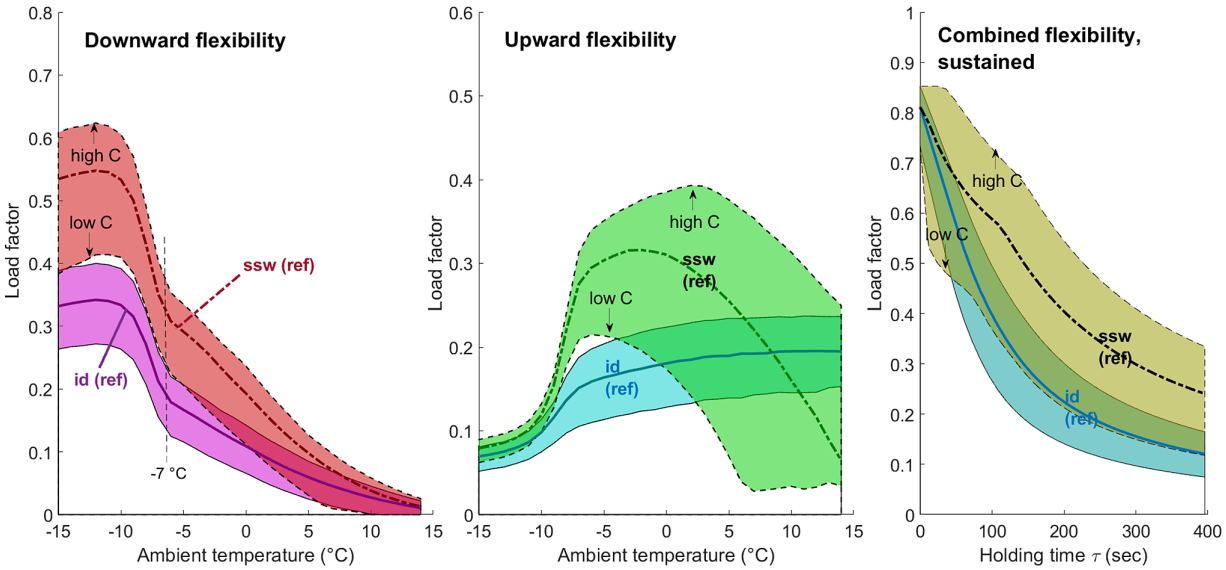

Figure 13. Regulation power in a bounded area of heat capacity, comparing ID with SSW. Downward (left) and upward (middle) power capacity sustained for $\bar{\tau} = 400$s depending on ambient temperature. Right: combined power capacity depending on holding durations.

displayed as a function of ambient temperature, whereas in the right image, the total power (up and down added; $\mathscr{F}^{+}(t,\tau)+\mathscr{F}^{-}(t,\tau)\leq 1\,\forall\, t,\tau,T_{\mathrm{amb}}$) is shown for varying holding duration, where outside temperatures have been sampled randomly and uniformly in the range $[-15,14]^{\circ}$C.

Often, but not always, significantly (20–50%) higher load deviations and longer holding times are observed for SSW compared to ID control. The SSW results are also more sensitive to the heat capacity C. Lacking a benchmark how much flexibility is really achievable, we can only state that the bounds for ID are stricter (more conservative) than those for SSW.

5.1.3. Results and Discussion

To interpret the results, we note that several input factors interact and contribute nonlinearly to the sustained flexibility. Basically, downward capacity decreases with the baseline, i.e., is less at milder outside temperatures under heating, whereas upward capacity increases. Due to the dynamic COP, the power capacity curve bends near $-7\,^{\circ}$C and assumes a maximum near $-12\,^{\circ}$C, below which the COP settles to one. Downward capacity decreases more steeply than upward capacity increases. Furthermore, the SSW upward capacity does not continue rising but assumes a maximum, unlike ID. We see a side effect of lockout constraints captured in the SSW but not in our ID estimate: Under mild ambient temperatures above $10\,^{\circ}$C, a decreasing heat load and a synchronously increasing COP shorten the heating cycles to the point where an absolute lockout duration of only 90 s blocks many opportunities for up regulation. The maxima positions and the variability of output depend on further interacting parameters, e.g., the thermostat band width D.

Regarding the actuation types, ID units that hold their operating states individually face a stronger constraint than SSW populations that approximate a target load state collectively; this indicates that more flexible energy might be extracted under SSW. On the contrary, ID control coordinates subgroups of appliances by time-relayed activation, which is not exploited for SSW control, and our ID estimate (Proposition 1) ignores the adverse lockout constraints.

Holding time characterizations are simpler to develop for SSW populations than for ID: they require only mean heating or cooling rates and no duty cycle distributions, and they treat service and recovery alike with flexibility signs just reversed. In contrast, ID/DLC methods are easier to implement with smart thermostats, whereas SSW requires randomization extensions "deep-down" in the HVAC thermostat controls.

5.2. Sensitivity Analysis

We analyze the sensitivity of the model response $\mathscr{F}^{\pm}(t,\tau)$, which depends on the ambient temperature $T_{\mathrm{amb}}(t)$, with respect to six input factors: a) Specific heat capacity (light to heavy construction); b) Heat transfer through the envelope (small to large); c) Heat consumption and dissipation of stored liquid or gaseous media (e.g., air change rate, infiltration); d) Building purpose (residential/non-residential); e) User tolerance (narrow to wide comfort interval); and f) Switching constraints (lockout duration).

5.2.1. Description

We carry out a local sensitivity analysis around the reference configuration that represents a mixed building stock with the parameters listed in Table A.1 and Table 2. For any input factor, the result value is strictly increasing or strictly decreasing at the reference point. Population parameters perturbed separately for each factor about the reference and denoted as high and low, respectively, are shown in **Table 3**. Modifications are carried out one-factor-at-a-time. All scenarios share the same HVAC technology (ASHP) and the same distribution of heating power. Two variations regarding building insulation (b) in **Figure 14**a and comfort tolerance (e) in Figure 14b, are visualized by error bars around the reference values. Similar as in Figure 13, results in the left diagram vary with ambient temperature and in the right diagram with holding duration. The upper diagrams show ID and the lower ones SSW populations.

Sensitivity indices have been estimated numerically for the reference configuration and are shown in **Table 4**. The four table rows distinguish down and up regulation as well as ID and SSW populations. The columns refer to the six input factors (a–f) and, as a seventh parameter, the std σ_W of thermal diffusion.[72] Each table entry is an average of 100 populations with 10 000 units sampled from the reference, low, and high configurations. These operate under uniform random temperatures in $[-15,14]^{\circ}$C. As there is no metric to compare qualitatively different influences, we empirically choose the values for low and high in Table 3 so to represent comparable perturbations in order to calculate sensitivity values.

5.2.2. Results and Discussion

Lockout duration (f) for ID units and diffusion σ_W in column 7 for ID and SSW can be discarded as noninfluential parameters, because they do not causally contribute to the estimates. Regarding prioritization of input factors, tightening or relaxing the thermostat band (e) has the biggest influence on flexible power and on holding duration. Heat capacity (a) and, with opposite signs, heat transfer coefficients (b) and—to a lesser extent—air change rates (c) have the next largest effects. The composition of the building stock by residential and non-residential buildings (d), which mainly differ in the area to volume ratios, is least influential. The impacts of heat transfer (b) and (c) vary between downward and upward flexibility, with the actuation type ID versus SSW, and with the outside temperature. For instance, heat

Table 3. Parameter perturbation in sensitivity experiments.

Building Scenario	Heat transfer $\overline{U}$	Air change Rate $\overline{r}_{\mathrm{ach}}$	Specific heat Capacity $\overline{C}$	Comfort Tolerance	R/NR Fractions
Ref configuration	$\mathscr{H}_{[0.1,1.3]}$	$\mathscr{H}_{[0.2,0.4]}$	$\mathcal{N}^{T}(0.85,0.15)$	$U_{[6,10]}$	0.75/0.25
Low configuration	$\mathscr{H}_{[0.1,0.7]}$	$\mathscr{H}_{[0.02,0.2]}$	$\mathcal{N}^{T}(0.6,0.1)$	$U_{[2,5]}$	0.25/0.75
High configuration	$\mathscr{H}_{[0.2,1.8]}$	$\mathscr{H}_{[0.4,0.8]}$	$\mathcal{N}^{T}(1.1,0.2)$	$U_{[6,10]}$	0.99/0.01

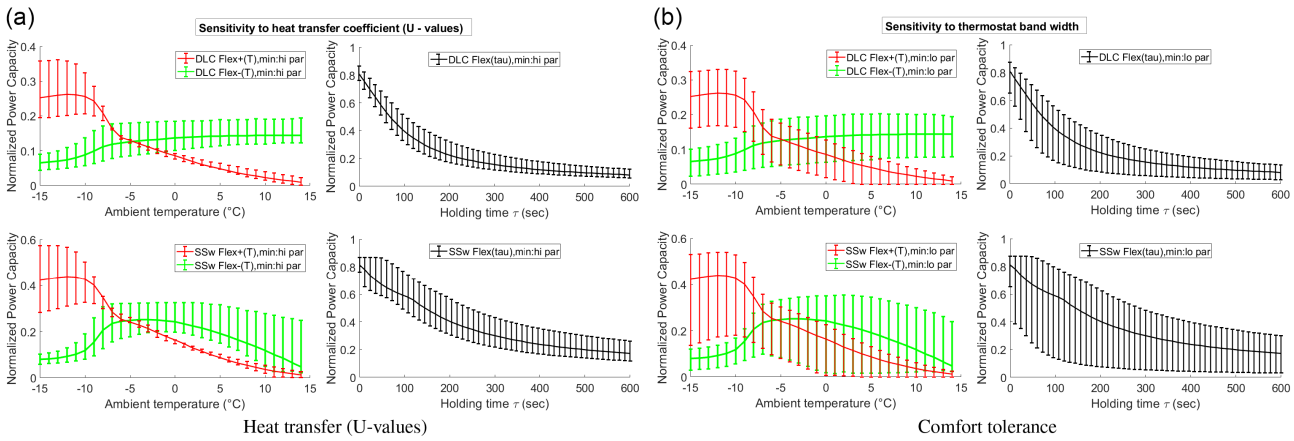

Figure 14. Sensitivity of flexible power capacity with regard to two dominant input factors.

Table 4. Sensitivity indices computed at the reference configuration; color intensity signifies magnitude.

	C	U	r_{ach}	R/NR	Comfort tol.	Lockout	dW
$\mathcal{F}^{-}_{ID}$	0.4621	−0.2573	0.4020	−0.2303	0.6426	−0.0006	0.0066
$\mathcal{F}^{+}_{ID}$	0.2557	−0.3257	0.0280	−0.0609	0.4979	0.0001	0.0259
$\mathcal{F}^{-}_{SSW}$	0.4623	−0.2242	0.4906	−0.2291	0.6319	−0.2816	0.0111
$\mathcal{F}^{+}_{SSW}$	0.5102	−0.5765	0.2945	−0.1979	0.8197	−0.4980	0.0372

transfer (b) affects downward regulation most under extreme cold, while infiltration (c) has a more uniform effect. In general, downward flexibility varies the more the higher the baseline load is, and upward flexibility the more the lower the baseline.

5.3. Realizability and Accuracy Tests

In this section, we evaluate how far the reduction of complex building units to TCL surrogates affects the aggregate level. We focus on the baseline HVAC consumption as the reference point of all flexibility, which depends mainly on the outside temperature, and compare the TCL-derived baseline estimates with an independent degree-day method to estimate the HVAC intensity (active power factor, APF, modified after the study by Gils[73]). Second, we compare the flexibility bounds of power, holding times (energy), and ramp rates, that are actually achievable by different TCL control algorithms in simulations, with the theoretical bounds, Equation (20) for $AM^{(0)}$, (35) for $AM^{(1)}$, and (33) for the ramping rate.

5.3.1. Description

Simulation: To achieve tolerable run times for up to $N \approx 10^5$ units, the Matlab simulation iterates a vectorized unit calculation structure in a common discrete-time loop. This is possible even though units have different constraints and control algorithms and their duty cycle durations range from one minute to several hours. A common time step dt always limits the overall temporal resolution compared with discrete-event simulation as in the study by Tindemans et al.[26] and, on the other hand, pretends a uniformly fast unit response, which is rather not achieved in practice by most heat pump controllers or refrigerators, for example. Readers may refer to the architecture diagram Figure B1 in Appendix Appendix B.

Active power factor: The independent APF method predicts an intensity factor of HVAC activity (ASHP, in our case) as a fraction of the total installed electricity. It requires 1) A description of a typical heating season in a region (duration, average ambient temperature, zero heating temperature, see **Table 5**); 2) Inventory data of the electricity (GW) installed in, and the floor space heated by, ASHP[74]; 3) The heating energy (GWh) consumed during a typical heating season by ASHP. Using the COP (A.2), an average electricity intensity factor per degree-hour at any outside temperature level is estimated. The active fraction of installed ASHP electricity is then interpolated on a temperature scale. These values are compared with the dynamic baseline estimate (3a).

We note that all methods, the TCL baseline formula, the TCL-based simulation, and the APF share, in one form or the other, the COP calculation method (A.2).

Table 5. Parameters of ASHP and heating season (values for Germany) to estimate the APF and the COP.

Symbol	Meaning	Unit	Value
–	Heating demand p.a. due to HP	TWh	11.8
P^{el}_{HP}	Installed electric HP power	GW_{el}	1.77
A_{HP}	Floor area heated by HP	m^2	8.1×10^6
	Heating Period:		
–	Average duration in days	d	225
T_{href}	Zero heating temperature	°C	17
$\overline{T}_{ah}$	Average ambient temperature	°C	3.5
η_{qfac}	Manufacturer efficiency factor		0.5
T_{fw}	ASHP supply temperature	°C	$U_{[35,50]}$
–	Minimum (design) temperature	°C	−15

Table 6. Baseline error [RMSE in %].

Configuration	N=25 000 $\sigma_W = 0.1$ $dt = 2s$	N=1000	N=100 000	$\sigma_W = 0.01$	$\sigma_W = 0.4$	$dt = 0.5s$	$dt = 5s$	Drift [% min^{-1}]
M-TCL	0.34	1.54	0.173	0.326	0.97	0.30	0.28	2.7
H-ASHP	0.28	1.42	0.165	0.331	1.31	0.36	0.31	5.6
C-ASHP	0.47	1.32	0.420	0.254	4.19	0.51	0.42	23.6

Formula-to-simulation comparison: The simulated baseline curve of an uncontrolled mixed-device (M-TCL) population in steady-state is compared with the dynamic baseline (3a) using the root mean squared error (RMSE in%), averaging 100 runs with random ambient temperatures. The following parametric influences on the errors are analyzed:

1) Composition by appliance types (M-TCL); 2) Population size N; 3) Thermal process parameters (diffusion σ_W, drift $\overline{\alpha}$); 4) Further "parasitic" influences, e.g., the time step dt.

In **Table 6**, the RMSE is listed for three population types running 10 000 s ($\approx$3 h) of simulated time. The accumulation of baseline errors is indicated in the rightmost column in% drift of the rated power per minute.

Formula-to-APF comparison: The baseline curves that include nonparticipating units in (5) with load factors of 1 (always on) or 0 (always off) are compared with the APF estimates over the outside temperature range shown in **Figure 15**. By varying also the parameters from Table 5, several curves result.

5.3.2. Results and Discussion

The largest baseline RMSE in Table 6 occurs in the space-cooling case (C-ASHP). Regarding population size, RMSE values decrease roughly proportional to $\sqrt{N}$. The error depends weakly on the level of thermal diffusion σ_W to the point where the process noise dominates the heating or cooling increments. As the calculated values after (3a) are biased due to small coefficient and approximation inaccuracies, their time-integrated deviations drift already in the short term. Drift is not an issue if the consumption of an existing population is measured to calibrate the baseline estimates. These can be adjusted to changing temperatures, e.g., (12).

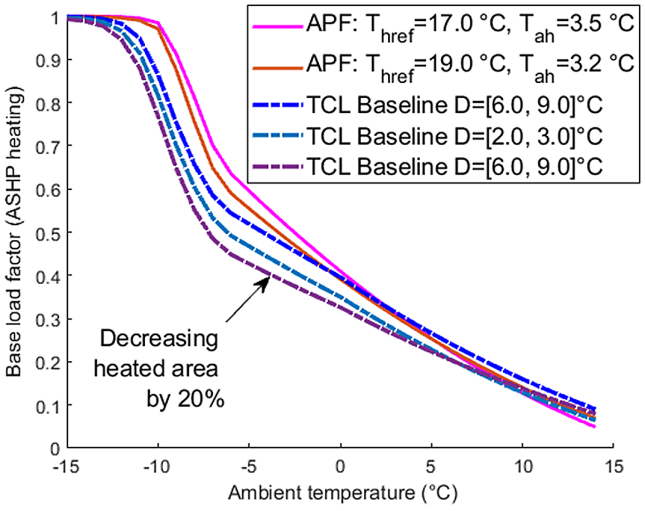

Figure 15. Baseline load depending on ambient temperature: APF (solid curves) compared with TCL formula (dashed curves).

Regarding the comparison with APF, we rate the agreement in Figure 15 as acceptable except below $-7\,^{\circ}$C where both methods become questionable. APF curves are truncated at 1 (100% load), because the ASHP no longer meet the full heat demand at $\approx -10\,^{\circ}$C with the inventory data assumed. In contrast, the calculated TCL baseline values remain $\approx$5% too low, although more than 95% of all units are already rated as "always on" (nonparticipating, baseline of one) at these very cold temperatures. Both the TCL and the APF formulas have adjusting screws so to achieve closer agreement. The two APF curves assume different zero heating and average seasonal temperatures. The TCL baseline curves differ by their assumptions of comfort tolerances and the total ASHP-heated floor area.

To test how far our bounds are realizable, we compare two SSW control algorithms (constant and linear hazard) with deterministic control of thermostat set points. At peak load deviations, the control algorithms tap only 30–50% of the constant energy bound (20) in simulations, while the mean bound (35) still predicts 75–90%. Measured ramp rates differ in the mean by less than 8% from the predictions, but individual samples vary considerably. For lack of space, we have outsourced the detailed experiment conditions and the graphical result into the Supporting Information (Section S-3.5.1).

5.4. Comparison with Ziras' Work

We compare our energy bounds with related parts of Ziras' work[39] where reference load signals are tracked under conditions of energy shortage. The authors' concerns were, among other topics, the oscillations following energy depletion.[39] They demonstrated that a simple stochastic controller oscillates less and dampens faster than PSC.

We focus on the points of energy depletion which indeed conform well to the TB bound from the study by Ziras et al.[39] and reproduce the conditions under which we obtain a similar performance. Using our locally linear SSW control algorithm with damping (Appendix B), we successfully suppress the controller-induced oscillations. More importantly, we explore and exacerbate the experiment conditions until more refined energy bounds than TB become essential, and analyze how energy sufficiency and tracking performance correlate.

5.4.1. Data Description

The TCL population studied by Ziras[39] (and earlier works) comprises synthetic devices specified by the reference values of thermal resistance $R_0 = 2\,(\mathrm{K\,kW^{-1}})$, capacitance $C_0 = 2\,(\mathrm{kWh\,K^{-1}})$, and energy conversion efficiency $\eta_0 = 3$. Thermostat band and outside temperature are fixed to $I = [T_-, T_+] = [22, 23]^\circ\mathrm{C}$, $T_a = 5\,^\circ\mathrm{C}$. The parameters are converted into TCL standard form (1) setting

$$T_0^\infty = T_a,\ T_1^\infty = T_a + \eta \cdot R \cdot P_n, \alpha = -\frac{1}{3600RC} \approx -6.94\mathrm{e}^{-5} \quad (39)$$

In the study by Ziras et al.,[39] parameter R alone is varied $\pm 10\%$ about R_0, which we tag as "homogeneous at degree $h = 0.9$," while $h < 1$ in general implies that all three parameters $X \in \{R, C, \eta\}$ are varied independently and uniformly about their reference values X_0: $X \sim U_{[X_0(1-h), X_0(1+h)]}$.

We specify the energy capacity Q^{TCL} as the number of seconds that the total rated power can be sustained ($\mathrm{W^{nre}s}$, normalized rated electricity [NRE]), which corresponds to

$$N \cdot P_n \cdot Q^{\mathrm{TCL}}/3600\,(\mathrm{kWh}) \quad (40)$$

in absolute values, i.e., $\approx 15.56 \cdot Q^{\mathrm{TCL}}$ assuming $N = 10\,000$ units with nominal power $P_n = 5.6\,(\mathrm{kW})$ each.

5.4.2. Experiment Description

For the regulation task, we assume the piecewise constant reference signal from the study by Ziras et al., [39] **Figure 16** therein, which demands roughly $600\,\mathrm{W^{nre}s}$ of flexible energy ($\bar{\gamma}^{\mathrm{bl}} \approx 0.52\,\mathrm{W^{nre}}$). The supply capacity assuming $\tau^{\mathrm{lck}} = 150\,\mathrm{s}$ lockout duration (smaller than the two values in:[39] $t^l_{\mathrm{off}} = 300\,\mathrm{s}, t^l_{\mathrm{on}} = 180\,\mathrm{s}$) is estimated to be $Q^{\mathrm{TCL}} \approx 341\,\mathrm{W^{nre}s}$ using (19) and is reasonably close to Ziras' estimate of $352\,\mathrm{W^{nre}s}$.[75]

By visual appearance, in **Figure 17**, the PSC and the stochastic controller both follow the reference perfectly well until, after $t \approx 700\,\mathrm{s}$, $175\,\mathrm{W^{nre}s} \mathrel{\widehat{=}} 50\%$ of the capacity have been used on the first upward load step $\Delta\bar{\gamma} \approx 0.26\,\mathrm{W^{nre}}$. Then the load starts oscillating about the baseline due to energy depletion. We can reproduce these results, which seem to confirm TB-like energy bounds, by assuming 1) The original population from the study by Ziras et al.[39] ($h = 0.9$), 2) No thermal diffusion ($\sigma_W = 0$), 3) An initial SoC of 0.5 or 50% ($T_0 \sim U_{[22,23]}$).

Using a constant-hazard SSW algorithm without damping, we reproduce similar oscillation behavior in **Figure 18**, except that our SSW controller slightly overshoots when the regulation sign changes. Tracking behavior of the locally linear algorithm with damping outlined in Appendix B is shown in **Figure 19**. Oscillations and overshooting have disappeared, but energy depletion obviously remains.

Prior tests with our ASHP and M-TCL populations—by default more heterogeneous and with diffusion—had however revealed a fundamentally different behavior when tracking the same energy-infeasible signal: 1) Control responses to load steps taper off quickly, 2) There are no oscillations to dampen, 3) The energy absorbed steadily declines with σ_W, 4) Accurate tracking near one or zero remains possible but requires an energy surplus beyond TB bounds.

To corroborate these findings, we vary the energy supply capacity of the population and the process noise σ_W as well as the reference trajectory to track, and measure the tracking accuracy (RMSE in%) and the retrieved energy fraction.[76] The energy supply grows by widening the thermostat band (in three energy steps: $[22, 23]$—lacking, $[21.5, 23.5]$—sufficient, and $[21, 24]^\circ\mathrm{C}$—surplus energy). Moreover, we scale the power amplitudes between 0.7 and 1.9 of their original values and accordingly scale the time to preserve the (TB) energy demand.

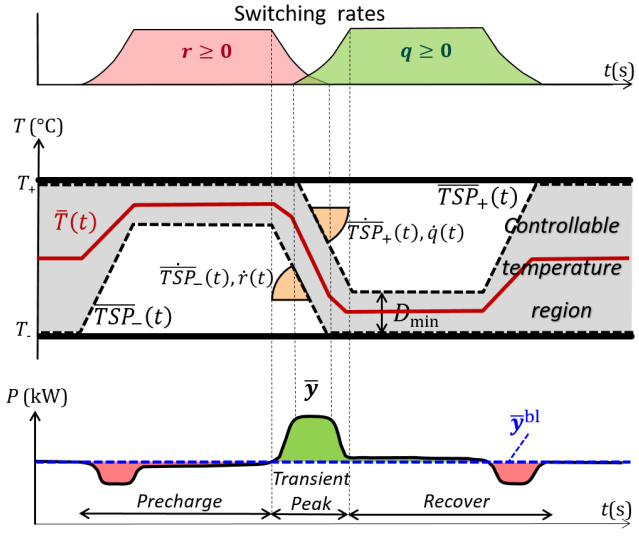

Figure 16. TCL population as a zero-order aggregation model AM$^{(0)}$: the mean temperature (middle) instantly responds to the switching rate profiles (top), and the aggregated load responds to mean temperature changes (bottom, assuming AC devices).

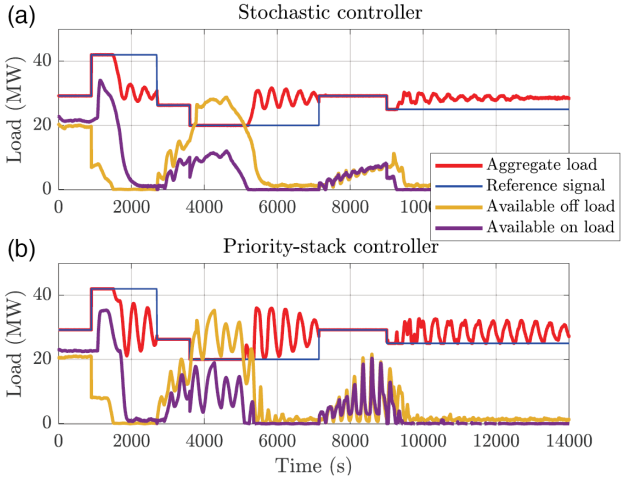

Figure 17. Tracking performance and evolution of the locked on and off loads. Reproduced with permission.[39] Copyright 2018, from the author, Ch. Ziras, Figure 6 therein.

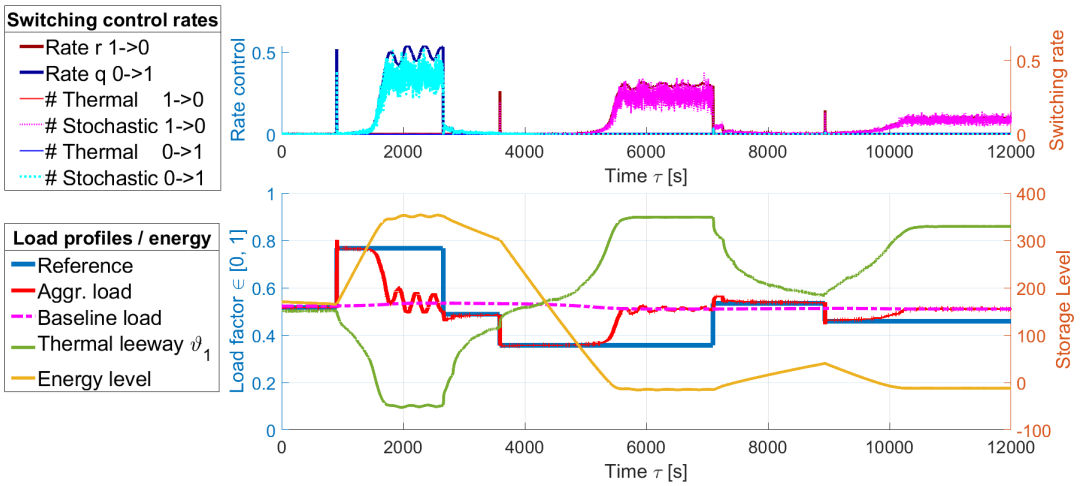

Figure 18. Tracking a piecewise constant reference load with Ziras' population ($h = 0.9$, no noise), using constant switching hazard.

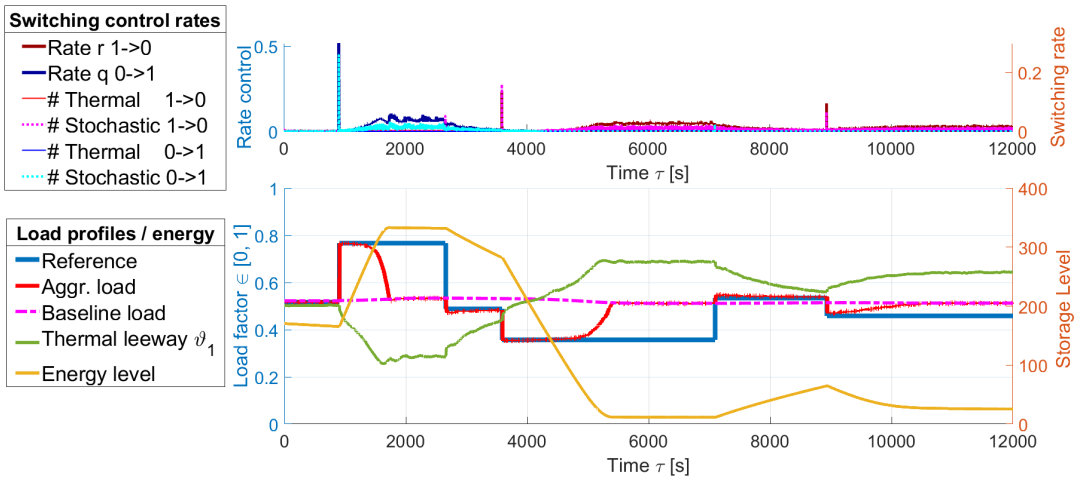

Figure 19. Tracking the reference load from Figure 18 using linear switching hazard and damping.

Acceptable RMSE values ($< 1\%$) are obtained only with surplus energy, low diffusion $\sigma_W \leq 0.03$, and scaling factors below 1.3.

5.4.3. Results and Discussion

Our SSW control with locally linear switching hazard shows good tracking performance, but for heterogeneous populations, it does only when excess energy w.r.t. TB capacity bounds is provided. Oscillations can be effectively suppressed by monitoring the energy level in the GBM.

For the regulation signals in Figure 18, the tracking error and the retrieved energy deteriorate with the following input factors, in decreasing order: 1) Thermal diffusion σ_W 2) Load amplitude (scaling factor) 3) Parameter heterogeneity (factor h).

We regard Ziras' population parameters[39] to be rather atypical of mixed residential HVAC devices and consider the test reference signal as benign. In the following section, we investigate more aggressive signals, where the holding-time-at-power level becomes even more important.

5.5. PJM Case Study

We aim to show that a refined characterization of flexible energy involving load levels (such as (35) and Proposition 4 in Section 4.3) is essential and useful. We let strongly heterogeneous TCL aggregations with thermal diffusion track the PJM dynamic regulation signals (RegD)[7] and assess their performance by varying load amplitudes and timing dynamics. Especially, we show that energy depletion at high power deviation levels and large tracking errors are highly correlated. DR resources may fail even if they provide sufficient energy by thermal-battery standards.

Reducing the load factors and increasing the rated power improves the performance and possibly the service rewards, but requires a higher reserve capacity to be compensated. We propose a simple design procedure to determine the necessary participation.

All experiments in this section are based on detailed TCL simulations, using SSW load control with constant or with locally linear switching hazard.

5.5.1. Mapping the PJM Signals

PJM interconnection electricity markets[7]), a major USA independent system operator, has been regularly publishing time-series data for load balancing since 2012. An area control error (ACE) power signal is split into two parts, a slower, energy-intensive RegA and a faster-responding RegD signal. The latter transmits a normalized value in $[-1, 1]$ every 2 s and is challenging to track, because aggressive switching between extreme levels causes long and steep-ramped (dis-)charging cycles, and also because the values often stay pegged at 1 or -1 for some time, i.e., require long holding durations. In reality, the values ± 1 represent a certain effective power imbalance or ACE of magnitude $\pm P^{\mathrm{reg}}(\mathrm{MW})$. In contrast, a maximal load factor of 1 at a DR resource uses an aggregated power of $P^{\mathrm{TCL}}(\mathrm{MW})$. The neutral RegD value zero is mapped onto the baseline load $\bar{y}^{\mathrm{bl}}$, which varies in time but is assumed constant, at least predictable, over a short service period of at most 2 h.

To carry out real load-balancing services, absolute power figures matter and must be related. Several options exist to adapt the capacity of a DR asset to the needs: a) Power scaling: The DR resource chooses a fraction $L \leq 1$ of the imbalance P^{reg}; i.e., it tracks the same signal shape according to its own capacity. This appears to be a necessary prior step in any case;[77–79] b) Time scaling: The DR resource follows a RegD signal in fast or in slow motion using a time factor b, which scales the energy demand as well: this is applicable only in simulation experiments; c) Load factor scaling: All DR units track a RegD signal scaled down in amplitude and use a fraction $\zeta \leq 1$ of their own power range. The aggregation is accordingly increased so to preserve power and energy contributions to the grid. Such part-load strategy is applicable in reality and is simulated below. While (a) defines the overall power and energy contribution, (c) decides the internal division of work. In Supporting Information, we document how to exactly match demand and supply of flexible power and energy and how to map the RegD signals using the factors L, b, and ζ.

Load-Scaling Simulations: As we ignore the power dimension by normalization, we must exercise caution when simulating option (c), and illustrate three ways of doing this.

Default: Scaling down the signal amplitude by ζ proportionally reduces the energy demand. Thus, the simulation pretends a better tracking performance than real.

Energy invariance: Scaling down the amplitude and stretching the time accordingly (factor b) keeps the energy demand constant but mitigates the PJM signal dynamics in slow motion.

Time acceleration: Preserving the amplitudes and compressing the time compensates an existing energy supply deficit and allows to test the power limits, but exacerbates the dynamics.

Lockout: Time acceleration becomes incompatible with any absolute duration of lockout since, eventually, all signal frequencies become too fast for load control. We exclude lockout from simulation in fast-motion.

5.5.2. Experiment Description

Illustrative example: In **Figure 20**, an M-TCL population ($N = 25000, \sigma_W = 0.08$) with 60% heating and 40% cooling appliances tracks a PJM RegD signal for 4 h starting from midnight December 3, 2018. The RegD energy demand of $Q^{\mathrm{reg}}_{\mathrm{max}} \approx 724\,\mathrm{W^{nre}s}$, mostly upward, exceeds the TCL supply $Q^{\mathrm{TCL}}_{0.5} \approx 170\,\mathrm{W^{nre}s}$ by more than four. Therefore, the PJM signal is time-compressed beforehand with $b = 4$. Tracking errors are measured in two ways: as root mean-squared error (RMSE) and by PJM precision score (see Section 4.5.6 in the operation manual[7]). The panel with four diagrams, from bottom (i) to top (iv), shows i) The TCL cannot hold load levels near 0 and 1 with $b = 4, \zeta = 1$; responses wear off toward the baseline. ii) They track a down-scaled (to 75%) signal with the same energy demand quite well ($b = 3,\ \zeta = 0.75$). iii) Time compression to cancel the realistic energy supply deficit ($b = 16,\ \zeta = 1$) allows levels 0 and 1 to be reached, albeit with reduced accuracy compared with (ii). Neither power capacity limits[37,39] nor lockout durations (not present) create a bottleneck. iv) With only 120 s lockout duration, tracking fails despite lowered amplitudes ($\zeta = 0.6,\ b = 2.4$). The (moderately) accelerated RegD dynamics causes a ramping speed bottleneck GW/s, whereas power capacity (GW) and energy capacity (GWs) are sufficient.

To verify the last claim in (iv), we repeat the simulation in original PJM time, without acceleration. To make it possible, we boost the energy supply by artificially slowing down thermal drift values and by relaxing the thermal comfort beyond reasonable limits. Load factors near 0 or 1 can now be reached, but, due to the lockout, the control dynamics is sluggish and tracking performance is unacceptable (RMSE 11.6 %, precision 0.754).

Experiment series: To corroborate individual findings, we conduct test series and aggregate the results statistically:

RegD signals: One series is carried out using data between July 2, 2019 and August 8, 2019 (summer) and another with data between January 1, 2019 and February 2, 2019 (winter). Each series comprises 50 services—2 h of continuous tracking—with random starting times regularly distributed over 20 days.

TCL configuration: The main series uses M-TCL with diffusion $\sigma_W = 0.05$, a SoC $\vartheta_0 = 0.5$ and TB energy supply $Q^{\mathrm{TCL}} \approx 350\,\mathrm{W^{nre}s}$. A comparison experiment is conducted using H-ASHP, whole buildings with a heavy construction and heat pump heating at $T_{\mathrm{amb}} = -2\,^{\circ}\mathrm{C}$. These supply less energy: $Q^{\mathrm{TCL}} \approx 180\,\mathrm{W^{nre}s}$.

The control algorithm with locally linear switching hazard and no damping works in the main series. One comparison experiment uses constant rates. Population size is $N = 2000$ in the main series; $N = 25000$ and N=100 000 for comparison.

Load factors are scaled down from $\zeta = 1, \cdots, 0.3$ in steps of 0.1, and four options are applied: 1) Original time; 2) Energy-invariant through slow motion; 3) Compensation of realistic supply deficit through time acceleration; 4) Options 2 and 3 combined: for each value $\zeta \leq 1$, a signal with the same TB energy demand is created first by stretching the time, and any remaining (realistic) energy deficit is then eliminated by compressing the time.

Numerical results and explanation: On the summer data under option 1, while ζ decreases from 1 to 0.3, RMSE decrease from 8.3% to 0.85%, and precision improves from 0.84 to 0.94 (**Figure 21**). The errors under option 2 become only slightly higher (for medium scaling factors ζ) and the curve flatter, if the energy demand is maintained through slow motion. The time-accelerating option 3 blends into the option 1 curve when

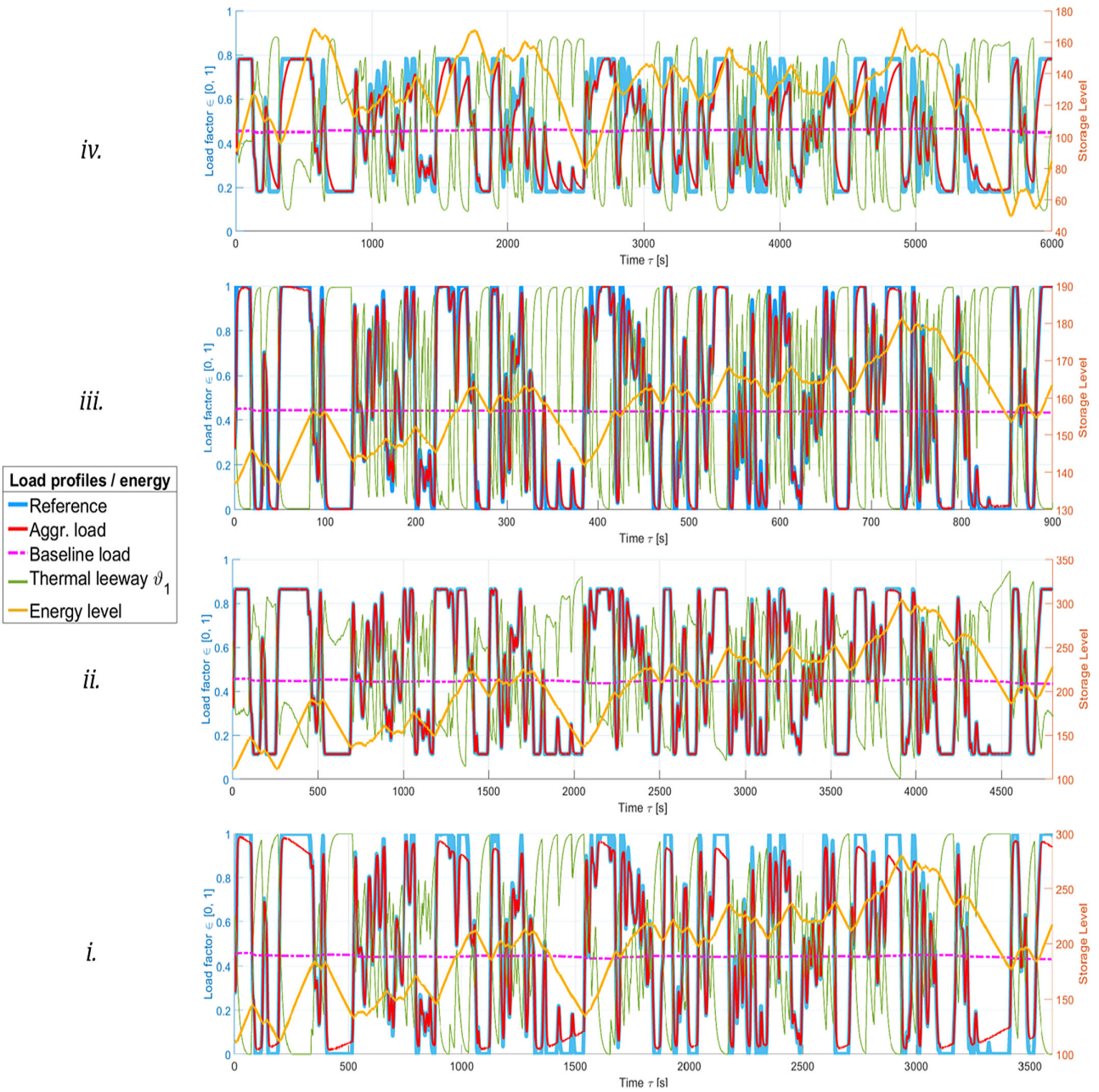

Figure 20. Tracking a RegD signal from December 3, 2018, 00:00:00 to 04:00:00 a.m. shown as a light-blue reference curve. From bottom to top (please note the different time axes and see text for illustration): i) full load range ($\zeta = 1, b = 4$): RMSE 4.96%, PJM precision score $s_p = 0.894$; ii) down-scaled ($\zeta = 0.75, b = 3$): RMSE 0.95%, $s_p = 0.981$; iii) time acceleration ($\zeta = 1, b = 16$): RMSE 3.87%, $s_p = 0.928$; iv) lockout 120s, ($\zeta = 0.6, b = 2.4$): RMSE6.43%$s_p = 0.811$.

amplitudes get small enough with $\zeta \to 0.3$ so that no (realistic) energy deficit remains. For $\zeta \approx 1$, option 3 does improve the errors compared with options 1 and 2, but the benefit is offset to a large part by the burden of fast dynamics (RMSE $\approx$ 5%). The combined option 4 indeed achieves the best of both worlds: RMSE $\approx$ 5% for $\zeta = 1$ and lowest RMSE 0.29% and highest precision values of 0.98 for $\zeta = 0.3$. For ζ near 1, option 4 behaves like 3 (little or no time stretching), whereas for small ζ, it blends into option 2: realistic energy equals TB energy, and no more deficit remains to compensate. With the winter data, we obtain similar results. H-ASHP (**Figure 22**, left) produces up to 20% higher RMSE than M-TCL for high load amplitudes ($\zeta > 0.7$), because of a smaller energy supply, respectively, a higher deficit, higher time acceleration factors are needed.

The tracking performance (RMSE, precision, energy absorbed) does not deteriorate with thermal diffusion as much on RegD signals as on the step signals considered in the previous Section 5.4.

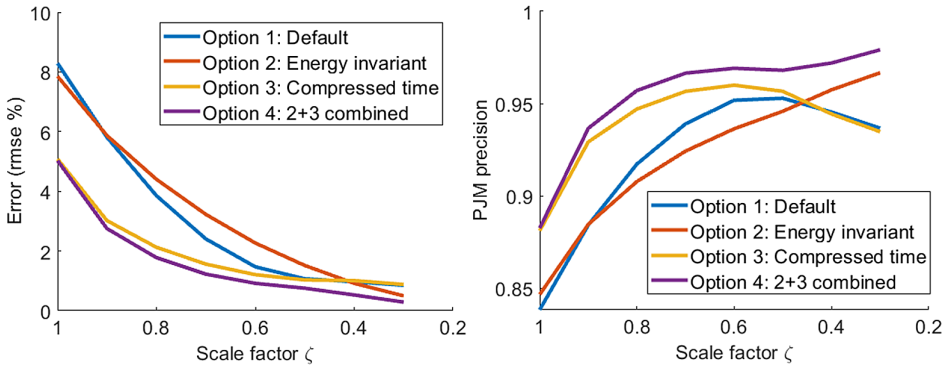

Figure 21. Tracking error (RMSE, left diagram) and precision (right diagram) in load scaling experiments on the summer RegD data.

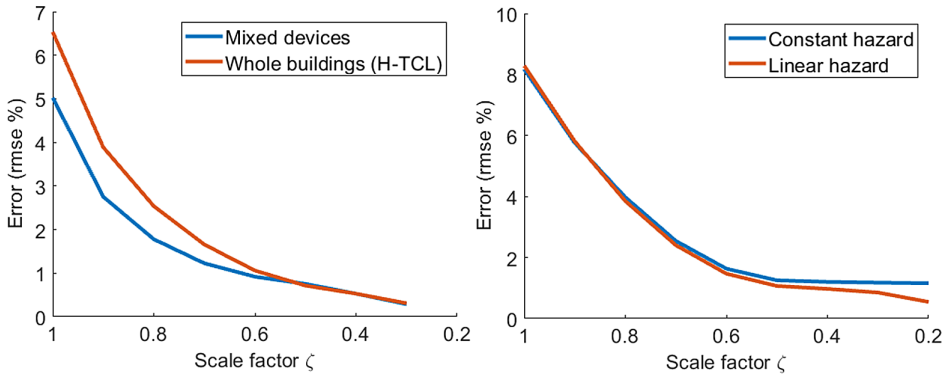

Figure 22. Tracking error (RMSE) using energy option 4 from Figure 21. Left: Comparing different populations (mixed devices against H-ASHP); right: comparing two control algorithms, linear hazard (L) against constant rate (C).

The control algorithms with constant (C) and with locally linear (L) switching achieve similar tracking precision, except that (C) performs slightly better on aggressive signals with fast excursions to and long stays on levels ± 1, and slightly worse on more moderate signals. (L) benefits from downscaling up to $\zeta = 0.2$ and achieves precision near 0.98 or RMSE of 0.28% (Figure 22, right).

However, performance scores alone poorly reflect the differences that become visible on a small scale (zoom-in diagram **Figure 23**). Due to the harder control force exerted, (C) produces fast and small oscillations (with periods < 1 s and amplitudes $<$ 1% of the load range) that are barely visible under (L). More severely, (C) needs 5–15 times more stochastic switches and, in particular, causes more fast compressor cycles than (L) to track the same signal; i.e., locally linear switching hazard does save equipment life. These effects are stronger for low $\sigma_W = 0.01$ than for high diffusion ($\sigma_W = 0.08$). In contrast, (L) affords fourfold-to-fivefold delays compared with (C). Response delays are estimated from the signal correlation as specified by PJM[7] and can be seen only in the zoomed view. Constant rate (C) follows the reference within one control cycle.

5.5.3. Scaling Heuristic

To find an appropriate scaling factor ζ, we can predict the realistic energy demand (RED) and compare it to the energy supply of a DR resource. Still, choices remain how to weigh extreme load factors. We propose an easy-to-use and practical heuristic to determine ζ, tailored specifically to PJM RegD signals and illustrated in **Figure 24**.

Both supply and demand of flexible energy are represented by durations (τ) over load deviations ($\Delta\gamma$). A demand curve for PJM signals (Figure 24, bottom) may be predicted from the load duration density and the service duration. The border maxima are due to the load levels near ± 1 that appear quite often in many RegD signals.

On the supply side, extreme positive and negative levels can be sustained only shortly but, with levels shrinking toward the baseline, the holding durations rise steeply (Figure 24 top). The supply curve results from (35) or (38) and resembles in shape Figure 2 rotated 90°.

To determine ζ, we compress the demand curve horizontally about the baseline until the border maxima fit under the supply curve. The durations are not accordingly increased, because it is more aggregated power that provides the energy. Due to the RegD shapes, the entire demand curve will fit under the supply curve once the boundary maxima do. The scaling problem will therefore have a unique solution for this type of regulation signals.

5.5.4. Summary of Results

The experiments show that energy-at-load-level is a strong indicator of tracking ability and performance. These connections are

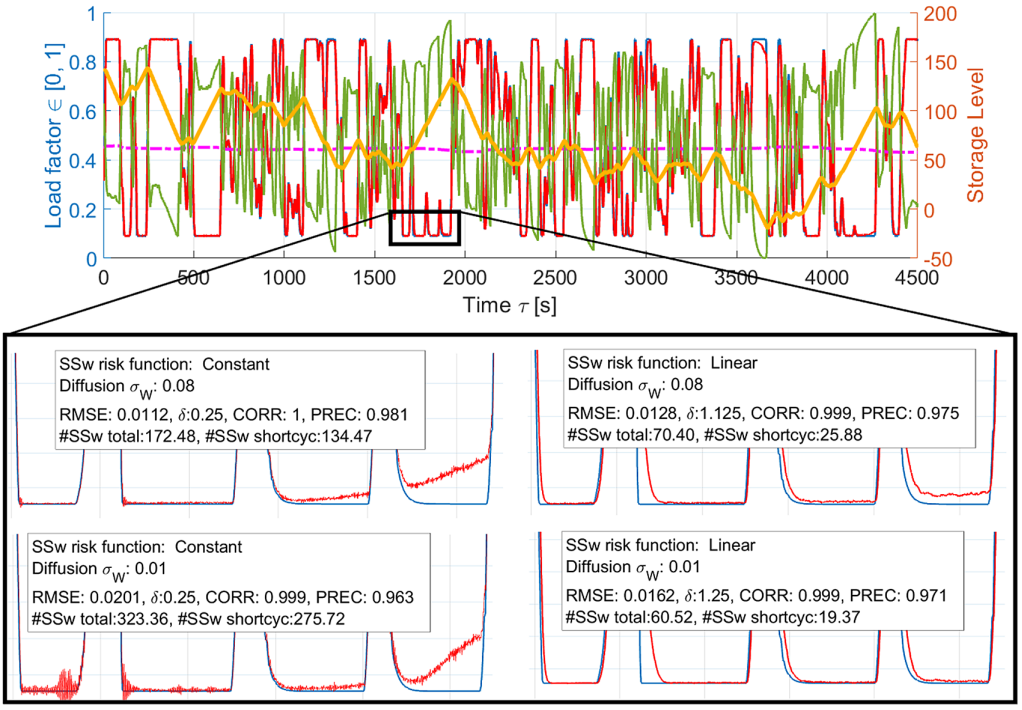

Figure 23. Comparison of constant (C) and locally linear (L) SSW algorithms using the data from Figure 20. The detailed tracking behavior is shown for the short time interval zoomed in (left-(L), right: (C), bottom-small, top-large σ_W).

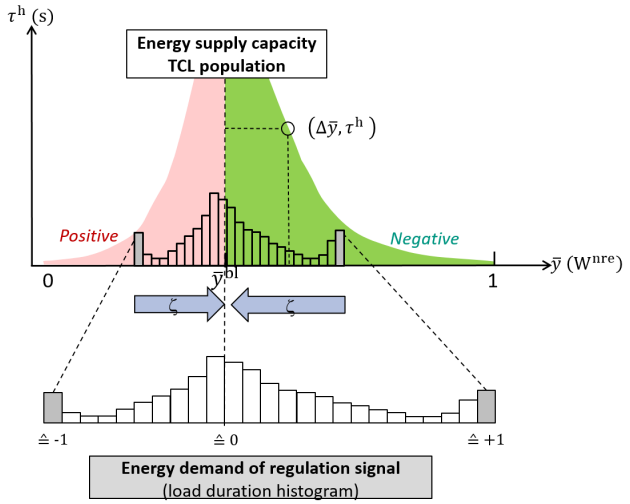

Figure 24. Matching demand and supply of flexible energy through scaling of power amplitudes $\zeta \leq 1$. The demand is represented by a RegD duration histogram (bottom), and the supply by a holding time curve of a TCL population (top).

not explained by pure power or pure energy bounds considered in the TB literature. For PJM dynamic regulation signals with a tendency towards peak amplitudes, the load factors can be reduced by participating more consumers. Another trade-off between response speed and compressor stress is made in the SSW control algorithm, e.g., constant versus locally-linear rates.

Using the refined energy characterization for planning and dimensioning of fast regulation services, DR designers may trade service quality against resource economy. In a specific regulatory framework, for example, aggregators that bid on an AS market and get paid for performance while they remunerate consumers through contractual incentives can roughly balance costs and revenues without solving a number of complex optimal scheduling problems.

6. Conclusions

This article addresses an urgent need and a scientific gap between the potential analysis and the operational scheduling of pooled flexible HVAC electricity demand. While the former is concerned about availability, feasibility, and revenues of DR at a large scale, the latter focuses on realizing aggregated flexibility in a specific regulatory and economic framework.

TCL aggregations in this work serve as surrogate models to characterize the flexibility performance ex ante through closed-form bounds, functions, or tables parametrized by operating conditions such as ambient temperature and load-switching constraints. The ability to estimate these bounds could as well be influencing the way efficiency interventions and local renewable energy systems (e.g., buildings with onsite generation) are planned and designed. A related outcome to be reported is the link to and the evaluation of energy efficiency measures planned to decarbonize the heat demand in the building stock, which are part of the energy transition.

Our focus is on fast regulation services at time scales between seconds and a few hours, and on dynamic performance. We derive new bounds on flexible power and holding times,

equivalently on payback power and recovery times, service frequency, and ramp rates. We refine thermal-battery-like energy bounds and obtain holding duration distributions that yield the retrievable energy at a given load difference with statistical guarantees. A case study using dynamic PJM regulation signals shows that energy at peak load deviations has a key impact on load tracking accuracy, a connection not explained in the literature before. We show how to improve the DR by reducing the power amplitudes and participating more loads.

Using survival processes, we develop new closed formulas of the expected switching temperatures in stochastically actuated TCL aggregations. Hazard functions have proved useful in defining a minimalist ODE formulation of population temperature and load dynamics, which is more transparent than black-box approaches and has interpretable coefficients. The resulting GBM allows to characterize flexible energy, power, and ramping ability of large aggregations.

We derive coarse estimates of flexibility criteria using closed-form transformations of input parameter densities, such as the HDD as a function of thermal constraints in Section 4.3.2, the (sustained) flexible power as a function of building stock parameters exemplified in Appendix Appendix A (**Table A1**) and in Section 5.1, and finally the density of nominal powers as an independent random input variable. This statistical aggregation approach handles parameter heterogeneity more directly than a clustering approach, which requires first creating and identifying numerous building-level models to group those with similar parameters into clusters.

Imminent applications of this work are seen in two fields: 1) Scenario analysis: Providing performance values of the DR side to assess the relative merits of different flexibility technologies; creating fact sheets for FS requirements specification, procurement, and trading; 2) Operational tool for aggregators: answering the feasibility question before bidding into an FS market and before scheduling the appliances; designing the services, e.g., deciding how many loads to participate.

Model identification details in the context of the present and further GBMs will be discussed in future work. There are several avenues of future research. The errors on the aggregation level when approximating continuously controlled units with several goal variables by simple on–off-controlled units should be better quantified, using higher-order unit models for comparison.[80] Furthermore, the flexibility performance bounds should be validated in DR field tests including large building ensembles.

Appendix A. Building Model Generation

To model individual buildings as parts of district or city-level aggregations, we start from a simplified first-order heat ODE; i.e., a single thermal mass lumps the building structure and the zones. The building ODEs are transformed into the standard TCL form (1) using the method in the study by Kohlhepp and Hagenmeyer[55] (Supporting Information, Equation (11) and (12) therein) which we briefly summarize. The first-order building ODE has several external driving forces. 1) Solar and sky radiation and equipment or occupancy gains are transformed into effective temperatures using the sol–air temperature concept;[81] 2) Conductive or convective heat transfer over the building envelope is proportional to the temperature difference between thermal storage and ambient air; 3) Available heating or cooling power is switched on or off depending on the storage temperature and turns the heat equation into a hybrid state description.

The resulting absolute heat flows are still unknown and are further decomposed using normalized quantities such as specific heat capacity, heat transfer coefficients, and design heat load intensity. Rules for building design and HVAC dimensioning provide the missing information and are combined with ratios between floor areas, thermal masses, and envelope surface areas that characterize certain building types and age classes; see the studies by Kohlhepp and Hagenmeyer; Appelhans et al.; Arendt; and Kemna and Acedo[55,74,82,83] and further references therein. These parameters characterizing the built environment are tagged "BE" in Table A1. We end up with four steering parameters for model generation: specific heat capacity $\overline{C}$, heat transfer coefficient $\overline{U}$, design heat load intensity $\overline{P}^{\mathrm{dli}}$, and air change rate r_{ach}. The conversions into TCL parameters are summarized in the following equations.

$$\alpha = -\frac{\overline{U} \cdot r_{\mathrm{S:V}}}{\overline{C} \cdot r_{\mathrm{MB:V}}} \tag{A.1}$$

Table A1. Building/HVAC stock and TCL model parameters.

Parameter	Meaning	Type	Unit	Value Range
$\overline{C}$	Specific heat capacity of thermal mass	TB	$\mathrm{kJ\,kg^{-1}\,K^{-1}}$	$f_C = \mathcal{N}^T(0.9, 0.2)$
$\overline{U}$	Heat transfer coefficient of envelopes	TB	$\mathrm{W\,m^{-2}\,K^{-1}}$	$f_U \in \mathscr{H}_{[0.1,1.3]}$
r_{ach}	Air change rate (ventilation and infiltration)	TB	$\mathrm{h^{-1}}$	$f_{\mathrm{Air}} \in \mathscr{H}_{[0.2,1.0]}$
$\overline{P}^{\mathrm{dli}}$	Design heat load intensity (HP, electrical)	HVAC	$\mathrm{W\,m^{-2}}$	$f_{\mathrm{Pdli}} \in \mathscr{H}_{[40,100]}$
η_{ASHP}	ASHP efficiency (COP)	HVAC	–	Computed from T_{amb}
T_{amb}	Ambient temperature	EN	°C	Measured, TRY
ΔT_{gl}	Temperature correction for gains and losses	EN	K	$\mathcal{N}(2.0, 1.0)$
$r_{\mathrm{V:A}}$	Gross building volume (GBV) per floor area	BE	$\mathrm{m^3\,m^{-2}}$	2.73(R) ⋯ 4.33(NR)
$r_{\mathrm{S:V}}$	Envelope surface area per volume (GBV)	BE	$\mathrm{m^2\,m^{-3}}$	0.51(R) ⋯ 0.32(NR)
$r_{\mathrm{MB:V}}$	Thermal mass per volume (GBV)	BE	$\mathrm{kg\,m^{-3}}$	528(R) ⋯ 400(NR)

$$T_0^{\infty} = T_{\text{amb}} + \Delta T_{\text{gl}} \tag{A.2}$$

$$T_1^{\infty} = T_0^{\infty} + \frac{\overline{P}}{(\overline{U} \cdot r_{\text{S:V}} + 0.34 r_{\text{ach}}) r_{\text{V:A}}} \tag{A.3}$$

$$\overline{P} = \eta_{\text{ASHP}}(T_{\text{amb}}) \cdot \frac{P_{\text{HP}}^{\text{el}}}{A_{\text{HP}}} \cdot \frac{\overline{P}^{\text{dli}}}{\mu_{f_{\text{Pdli}}}} \tag{A.4}$$

where $\overline{U} \sim f_U, \overline{C} \sim f_C$ in (A1), $r_{\text{ach}} \sim f_{\text{Air}}$ (A3), and $\overline{P}^{\text{dli}} \sim f_{\text{Pdli}}$ (A4) are sampled from distributions listed in Table A1.

Meaning of parameter types in column 3: TB—Thermophysical building parameters, HVAC—parameters describing the HVAC stock, EN—Environmental parameters (weather and occupancy), BE—structural parameters of the built environment. Values in column 5 are sampled from specific distributions (e.g., $\mathcal{N}^T(\cdot,\cdot)$: normal distribution truncated to positive values with given mean and standard deviation) or from a family of densities, e.g., $\mathscr{H}(\cdot,\cdot)$: histogram partitions on the given domain interval. For BE, only two class means are specified (R: residential, and NR: nonresidential buildings). Ambient temperatures are measured time series data (TRY: test reference year). Overlines symbolize normalized or specific quantities. Parameter values for specific experiments and thermal comfort bounds are defined separately in Section 5.1, e.g., Table A1.

For example, the thermal power $\overline{P}$ in the numerator counteracting the heat gains or losses in the denominator of (A3) together define the temperature span $|T_1^{\infty} - T_0^{\infty}|$ between active and passive states. Transmission losses stated per unit area of building envelope and ventilation losses per unit air volume are translated into the joint unit of floor area by exploiting the morphological relations $r_{\text{S:V}}, r_{\text{V:A}}$. The conversion factor $\rho_{\text{air}} \overline{C}_{\text{air}}/3600 \approx 0.34\,\text{W}\,\text{m}^{-3}\,\text{K}^{-1}$ is due to the specific heat and density of room air.

We consider air-source heat-pumps (ASHPs) for heating and cooling, which can operate during much of the year and consume rather much electricity available for DR in compressors and pumps.[84–86] η_{ASHP} in (A.4) connects the available thermal power $\overline{P}$ to the installed ASHP electricity per m^2, which is expressed as a distribution of design load intensities f_{Pdli} and covers many building types. Its mean is scaled using the totally installed HP electricity $P_{\text{HP}}^{\text{el}}$ and the HP-heated floor are A_{HP} in a given region (see the study by Appelhans et al.[74] for Germany). The ASHP efficiency is composed of the ideal $\eta_{\text{carnot}} > 1$ of a reversible Carnot process and a manufacturer quality factor $\eta_{\text{qfac}} \leq 1$ and varies with the outside temperature T_{amb}

$$\eta_{\text{ASHP}}(T_{\text{amb}}) = \eta_{\text{qfac}} \cdot \eta_{\text{carnot}} = \eta_{\text{qfac}} \cdot \frac{T_{\text{w}} + 273.15}{T_{\text{w}} - T_{\text{c}}} \tag{A.5}$$

where $T_{\text{w}} > T_{\text{c}}$ denote the warm and cool sides of heat exchange, respectively. In heating mode, $T_{\text{c}} = T_{\text{amb}}$ and T_{w} equals T_{fw}, the supply temperature of the heat distribution circuit. In cooling mode, $T_{\text{w}} = T_{\text{amb}}$ and $T_{\text{c}} = T_{\text{fw}}$, the supply point of cooled air.

The exact figures obtained using (A.5) depend on the definition of the reference temperatures, i.e., the ASHP system boundaries. We place them outermost while some texts determine T_{w} and T_{c} further inside, for instance as the mean heat exchanger temperatures between compressor and warm side, respectively, between evaporator and cool side. Narrowing in the system boundaries however lowers the Carnot efficiency. This is already accounted for in the quality factor $\eta_{\text{qfac}} \leq 1$ in (A.5), which models further efficiency losses due to auxiliary pumps and fans, heat exchangers, or pressure drops. We use (A.5) to model fictitious ASHP populations; therefore, evaporator and condenser temperatures are unknown.

Equation (A.5) models the COP above a threshold temperature near −7 °C reasonably well; below that, the COP keeps decreasing toward one, near −14 °C, blending into pure electric heating as an emergency operation state.[84]

Appendix B. Control Algorithm

A simple algorithm for testing the flexibility bounds and for actual reference tracking, based on the GBM in Section 4.3.1, is briefly outlined. **Figure B1** shows the communication and control architecture envisaged for SSW. To follow a reference load signal $\overline{y}^{\text{ref}}$, the aggregated load $\overline{y}^{\text{agg}}$ is measured and switching rates q, r are calculated that attempt to cancel the error $e(t) := \overline{y}^{\text{ref}}(t) - \overline{y}^{\text{agg}}(t)$ in one time step. We solve the derivative $\overline{y}^{\text{agg}}(t)$ (27) for the control variables that effect the desired load changes for small dt

$$\begin{aligned} e[k] &= \overline{y}^{\text{ref}}[k] - \overline{y}^{\text{agg}}[k] \approx \overline{y}^{\text{agg}}[k+1] - \overline{y}^{\text{agg}}[k] \\ &\approx \overline{y}^{\text{agg}}(q,r)[k] \cdot \text{d}t \end{aligned} \tag{B.1}$$

As both switching rates are positive, they take care of different signs of the tracking error

$$r[k] > 0 \overset{\text{Load}\downarrow}{\Leftrightarrow} e[k] < 0,\ q[k] > 0 \overset{\text{Load}\uparrow}{\Leftrightarrow} e[k] > 0 \tag{B.2}$$

We adopt a simple split-range controller with slight actuator redundancy; both rates can be nonzero. If $q[k-1] > 0$ but $e[k] <$

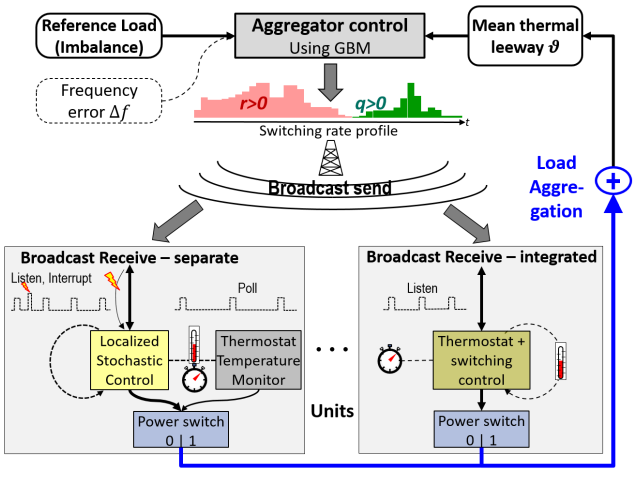

Figure B1. Control architecture for SSW with rate profiles broadcast and processed independently. Two options at unit level are outlined: integrated into a thermostat controller (right) or realized separately (left, e.g., through a programmable logic controller). The latter can respond to emergency events from the grid and implement local switching control but requires a shared access with the appliance thermostat controller.

0 and likewise, if $q[k-1] < 0$ but $e[k] > 0$, our algorithm may either decrease q or increase r to reduce the error. It reduces the nonzero before building up the zero rate.

TCL units adapt the broadcast rates q, r to their own needs (e.g., lockout, phase- or temperature-dependent switching hazard) based on local measurements. The appropriate control signals at aggregation level depend on the thermal leeway of the population, which is closely related to its state-of-charge as a TB. We define the normalized leeway-on ϑ_1 as

$$\overline{\vartheta}_1 := p^H \frac{\overline{T}_0^H - \overline{T}^H}{\overline{D}^H} + (1-p^H)\frac{\overline{T}^C - \overline{T}_0^C}{\overline{D}^C} \in [0,1] \tag{B.3}$$

where $\overline{T}^{\{H|C\}}$ denote the mean temperatures separately of AH, respectively, AC appliances; $\overline{T}_0^{\{H|C\}}$ the mean thermostat points where these must switch off and the leeway to switch on therefore vanishes, and $\overline{D}^{\{H|C\}}$ are normalization widths. Taking separate means of AH and AC devices assures positive signs for all terms and no cancellation. If the fraction $0 \le p^H \le 1$ of AH devices in the population is known—as contractual information or through dynamic log-in for service—both substates $\overline{T}^H, \overline{T}^C$ are observable from the measured load $\overline{y}^{\text{agg}}$[87]

$$\begin{pmatrix} \dot{\hat{T}}^H \\ \dot{\hat{T}}^C \\ \dot{\hat{y}} \end{pmatrix} = A \begin{pmatrix} \hat{T}^H \\ \hat{T}^C \\ \hat{y} \end{pmatrix} + B_0(\xi_1(t)) + B_u^{(L)}(\hat{T}^H, \hat{T}^C, \hat{y}) \begin{pmatrix} q \\ r \end{pmatrix} + L_{33}(\hat{y} - \overline{y}^{\text{agg}}) \tag{B.4}$$

L_{33} denotes the observer gain from a 3×3 diagonal observer matrix L. The mean temperatures $\overline{T}^H, \overline{T}^C$ are separately observable for AH and AC devices due to the nonzero coefficients A_{13}, A_{23}.

To prevent oscillation and overshooting in case of energy depletion while tracking an infeasible reference signal,[39] the control algorithm has an optional modulation factor $w_Q^{\text{ctrl}} \in [0,1]$ downstream in the control section, which lowers the rates q, r when running short of energy.[88]

The factor w_Q^{ctrl} is designed to approach zero if either one condition is satisfied: a) $Q(t)$ approaches the upper energy limit Q^{TCL} and $\overline{y}^{\text{bl}} \le \overline{y}^{\text{agg}} < \overline{y}^{\text{ref}} (q > 0, \dot{Q} \ge 0)$, or b) $Q(t)$ approaches the lower limit 0 and $\overline{y}^{\text{bl}} \ge \overline{y}^{\text{agg}} > \overline{y}^{\text{ref}} (r > 0, \dot{Q} \le 0)$.

The damping factor w_Q^{ctrl} swings back (using a low-pass filter) to its default value 1, otherwise.

$$w_Q^{\text{ctrl}} := \begin{cases} 1 + (\overline{Q}(t) - 1) \cdot w(t), & \text{if } \overline{y}^{\text{agg}}(t) > \overline{y}^{\text{ref}}(t) \\ 1 - \overline{Q}(t) \cdot w(t), & \text{if } \overline{y}^{\text{agg}}(t) < \overline{y}^{\text{ref}}(t) \end{cases}$$
$$\text{where } w(t) := \max\left\{0, \min\left\{1, \frac{\overline{y}^{\text{bl}}(t) - \overline{y}^{\text{ref}}(t)}{\overline{y}^{\text{agg}}(t) - \overline{y}^{\text{ref}}(t)}\right\}\right\} \in [0,1], \tag{B.5}$$
$$\text{and where } \overline{Q}(t) := \max\left\{0, \min\left\{1, \frac{Q(t)}{Q^{\text{TCL}}}\right\}\right\}$$

The damping option is used for comparative testing in Section 5.4.

We neither model nor measure temperature distributions to estimate the mean thermal leeway, nor do we maintain a central list sorted by urgency of switching such as in the PSC algorithm.[33] Unlike the study by Tindemans et al.,[26] our algorithm controls the load directly from the thermal leeway and not through the mean heating or cooling rates.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors cordially thank the reviewers for their insightful comments which improved the presentation and pointed out additional connections. The work herein has been funded by the Helmholtz Association in the Research Field Energy, Program Energy System Design.

Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Research data are not shared.

Keywords

ancillary services, demand responses, flexibility characterization, stochastic switching controls, thermostatically controlled loads

Received: March 29, 2021
Revised: May 31, 2021
Published online:

[1] D. Ürge-Vorsatz, L. F. Cabeza, S. Serrano, C. Barreneche, K. Petrichenko, *Renew. Sustain. Energy Rev.* **2015**, *41*, 85.
[2] J. Hu, G. Yang, K. Kok, Y. Xue, H. Bindner, *J. Mod. Power Syst. Clean Energy* **2017**, *5*, 451.
[3] F. L. Müller, B. Jansen, *Appl. Energy* **2019**, *239*, 836.
[4] W. Li, P. Xu, X. Lu, H. Wang, Zh. Pang, *Energy* **2016**, *114*, 981.
[5] P. Kohlhepp, H. Harb, H. Wolisz, S. Waczowicz, D. Müller, V. Hagenmeyer, *Renew. Sustain. Energy Rev.* **2019**, *10*, 527.
[6] Z. E. Lee, Q. Sun, Z. Ma, J. Wang, J. S. MacDonald, K. Max Zhang, *ASME J. Eng. Sustain. Bldgs. Cities* **2020**, *1*.
[7] PJM, *PJM Ancillary Services. Pennsylvania – New Jersey – Maryland interconnection electricity markets (PJM)*, PJM, Norristown, PA **2018**, 121, www.pjm.com/-/media/documents/manuals/m12-redline.ashx.
[8] E. Vrettos, F. Oldewurtel, G. Andersson, *IEEE Trans. Power Syst.* **2016**, *31*, 4272.
[9] E. Vrettos, Ch Ziras, G. Andersson, *IEEE Trans. Power Syst.* **2017**, *32*, 2924.
[10] S. Burger, J. P. Chaves-Ávila, C. Batlle, I. J. Pérez-Arriaga, *Renew. Sustain. Energy Rev.* **2017**, *77*, 395.
[11] M. Joos, I. Staffell, *Renew. Sustain. Energy Rev.* **2018**, *86*, 45.
[12] Ch Eid, P. Codani, Y. Perez, J. Reneses, R. Hakvoort, *Renew. Sustain. Energy Rev.* **2016**, *64*, 237.
[13] F. Oldewurtel, T. Borsche, M. Bucher, P. Fortenbacher, M. G. V. T. Haring, J. L. Mathieu, O. Megel, E. Vrettos,

G. Andersson. in *Proc. IREP Symp.: Bulk Power System Dynamics and Control. IREP 2013*, IREP (International Institute of Research and Educational in Power System Dynamics), Rethymnon **2013**, pp. 1– 24.
[14] K. O. Aduda, T. Labeodan, W. Zeiler, *Energy Bldgs* **2018**, *159*, 164.
[15] J. Zhang, A. D. Domnguez-Garca, *IEEE Trans. Smart Grid* **2017**, *9*, 4577.
[16] D. E. M. Bondy, *Ph.D. Thesis*, Technical University of Denmark, **2017**.
[17] J. Le Baut, G. Leclerq, G. Vigano, M. Z. Cegefa, Characterization of Flexibility Resources and Distribution Networks, Technical Report, EU/H2020 SmartNet Project, Deliverable 2.2, SINTEF Energi AS, **2017**, http://smartnet-project.eu/.
[18] We aim at performance data sheets to characterize DR resources and not entire power grids, although the benefit in each use case well depends on properties of the surrounding grid, such as topology and inertia.
[19] W. Cardinaels. Demand Response for Families, Technical Report, LINEAR Consortium, **2014**, 136, www.energyville.be/sites/energyville/files/downloads/2020/boekje_linear_okt_2014_boekje_web.pdf.
[20] K. Kessels, C. Kraan, L. Karg, S. Maggiore, P. Valkering, E. Laes, *Sustainability* **2016**, *8*, 929.
[21] N. H. S. Duong, P. Maillé, A. K. Ram, L. Toutain, *IEEE Trans. Smart Grid* **2019**, *10*, 1826.
[22] X. Xu, Q. Lyu, M. Qadrdan, J. Wu, *IEEE Trans. Sustain. Energy* **2020**, *11*, 2617.
[23] E. Dall Anese, S. Guggilam, A. Simonetto, Y. Ch. Chen, S. V. Dhople, *IEEE Trans. Power Syst.* **2017**, *33*, 1868.
[24] W. Zhang, J. Lian, C.-Y. Chang, K. Kalsi, *IEEE Trans. Power Syst.* **2013**, *28*, 4655.
[25] J. L. Mathieu, S. Koch, D. S. Callaway, *IEEE Trans. Power Syst.* **2013**, *28*, 430.
[26] S. H. Tindemans, V. Trovato, G. Strbac, *IEEE Trans. Control Syst. Technol.* **2015**, *23*, 1685.
[27] L. C. Totu, R. Wisniewski, J. Leth, *IEEE Trans. Control Syst. Technol.* **2017**, *25*, 1537.
[28] R. De Coninck, L. Helsen, *Appl. Energy* **2016**, *162*, 653.
[29] R. G. Junker, A. G. Azar, R. A. Lopes, K. B. Lindberg, G. Reynders, R. Relan, H. Madsen, *Appl. Energy* **2018**, *225*, 175.
[30] V. Trovato, S. H. Tindemans, G. Strbac, *IET Gen. Trans. Distribut.* **2016**, *10*, 585.
[31] V. Trovato, I. M. Sanz, B. Chaudhuri, G. Strbac, *IEEE Trans. Power Syst.* **2016**, *32*, 2106.
[32] M. Geidl, G. Koeppel, P. Favre-Perrod, B. Klöckl, G. Andersson, and K. Fröhlich. in *Third Annual Carnegie Mellon Conference on the Electricity Industry*, Vol. *13*, Mellon University Carnegie, Pittsburg, PA **2007**, p. 14.
[33] H. Hao, B. M. Sanandaji, K. Poolla, T. L. Vincent, *IEEE Trans. Power Syst.* **2014**, *30*, 189.
[34] H. Hao, B. M. Sanandaji, K. Poolla, T. L. Vincent, *Energy Policy* **2015**, *79*, 115.
[35] A. Ulbig. *Ph.D. Thesis,* ETH Zurich, **2014**.
[36] L. Zhao, W. Zhang, H. Hao, K. Kalsi, *IEEE Trans. Power Syst.* **2017**, *32*, 4721.
[37] B. M. Sanandaji, T. L. Vincent, K. Poolla, *IEEE Trans. Sustain. Energy* **2015**, *7*, 865.
[38] S. Khan, M. Shahzad, U. Habib, W. Gawlik, P. Palensky, in *IEEE Int. Conf. on Industrial Technology*, IEEE, Piscataway, NJ **2016**, pp. 570–575.
[39] Ch Ziras, S. You, H. W. Bindner, E. Vrettos, *Power Systems Computation Conf.* IEEE, Piscataway, NJ **2018**, pp. 1–7.
[40] A. Abiri-Jahromi, F. Bouffard, *IEEE Trans. Power Syst.* **2015**, *31*, 1972.
[41] A. Coffman, N. Cammardella, P. Barooah, S. Meyn. Aggregate capacity of TCLs with cycling constraints, arXiv preprint arXiv:1909.11497, **2019**.
[42] A. R. Coffman, N. Cammardella, P. Barooah, S. Meyn, in *American Control Conf.*, IEEE, Piscataway, NJ **2020**, pp. 527–532.
[43] L. Barth, N. Ludwig, E. Mengelkamp, Ph. Staudt, *Comput. Sci. Res. Dev.* **2018**, *33*, 13.
[44] B. Roossien. *Mathematical Quantification of Near Real-Time Flexibility for Smart Grids, Technical Report*, Energy Research Centre of the Netherlands (ECN), **2010**.
[45] R. Malhame, Ch.-Y. Chong, *IEEE Trans. Autom. Control* **1985**, *30*, 854.
[46] D. S. Callaway, *Energy Convers. Manage.* **2009**, *50*, 1389.
[47] D. Angeli, P.-A. Kountouriotis, *IEEE Trans. Control Syst. Technol.* **2012**, *20*, 581.
[48] S. Bashash, H. K. Fathy, *IEEE Trans. Control Syst. Technol.* **2013**, *21*, 1318.
[49] Ch Ziras, E. Vrettos, Y. Shi, *J. Mod. Power Syst. Clean Energy* **2017**, *5*, 43.
[50] M. S. Nazir, St. C. Ross, J. L. Mathieu, I. A. Hiskens, *IFAC-Papers Online,* **2017**, *50*, 8873.
[51] J. Hu, J. Cao, T. Yong, J. Guerrero, M. Chen, Y. Li, *IEEE Trans. Control Syst. Technol.* **2017**, *25*, 1586.
[52] V. Trovato, F. Teng, G. Strbac, *IEEE Trans. Smart Grid* **2018**, *9*, 5067.
[53] A. Kleidaras, A. E. Kiprakis, J. S. Thompson, *Energy* **2018**, *145*, 754.
[54] J. L. Mathieu, M. E. H. Dyson, D. S. Callaway, *Energy Policy* **2015**, *80*, 76.
[55] P. Kohlhepp, V. Hagenmeyer, *Energy Technol.* **2017**, *5*, 1084.
[56] J. Hu, J. Cao, M. Chen, J. Yu, J. Yao, S. Yang, T. Yong, *IEEE Trans. Power Syst.* **2016**, *32*, 3157.
[57] The FPE with a pure drift term is applied in the study by Tindemans et al.[26] Some works, the study by Zheng et al.,[89] use a FPE with diffusion term to simulate limited TCL heterogeneity.
[58] J. Ponocko, J. V. Milanovic, *IEEE Trans. Power Syst.* **2018**, *33*, 5446.
[59] In contrast, TB-like characterizations couple power and duration through a single energy capacity bound, while power capacity, the maximum instantaneous load difference, is the second parameter.
[60] For ease of notation, random variables are not capitalized. It is clear from the context whether a variable is deterministic or random or is a realization of a RV. If X is a RV distributed after a (not explicitly named, arbitrary) distribution that has a density fx, we also use the shorthand notation X~ fx.
[61] The TCL parameters $T_0^\infty, T_1^\infty, \alpha$ jointly depend on the outside, or ambient, temperature T_{amb}, on the power P^{TCL}, and on the heat gains or losses of storage units. The COP η also depends on T_{amb}. To see the connections, please refer to Figure 1 and 7 for details, where we show how to convert (first-order) building and HVAC models schematically into TCL using distributions of basic parameters that describe the stock.
[62] R. E. Mortensen, K. P. Haggerty, *IEEE Trans. Power Syst.* **1990**, *5*, 243.
[63] Traversing the entire band *I* back and forth arguably is the simplest, but not the only way to estimate the (mean) baseline.
[64] C. Perfumo, E. Kofman, J. H. Braslavsky, J. K. Ward, *Energy Convers. Manage.* **2012**, *55*, 36.
[65] A detailed derivation is provided in Supporting Information; as well for propositions 2–4 and Equation (12), (24), and (27), (29,30,32), and (35). Proofs are omitted in the following text without being pointed out each time.
[66] E. Webborn, R. S. MacKay, *Complexity* **2017**, *2017*, 26, Article ID 5031505, https://doi.org/10.1155/2017/5031505.

[67] D. R. Cox, D. O. Oakes, *Analysis of Survival Data*., Chapman & Hall Ltd, London **1984**, 212.
[68] However, load levels cannot be brought back by reverse actions, as Figure 6 suggests. For this reason, we need a refined model $AM^{(1)}$.
[69] The slopes κ assumed in (17) ($UM^{(0)}$) correspond to first-order quantities ($UM^{(1)}$) through $\kappa_s \approx \alpha(T_s - T_s^{\infty})$. In $AM^{(0)}$, the $\overline{TSP}$ and $\overline{T}$ curves follow the control input instantly; in $AM^{(1)}$, $\overline{T}$ dues its momentum to the mean drift rate $\overline{\alpha}$.
[70] B.C. Arnold, N. Balakrishnan, H.N. Nagaraja, *A First Course in Order Statistics*, Vol. *54,* SIAM, Philadelphia, PA **1992**.
[71] Note that p is the fraction of TCL that might not comply!
[72] Diffusion should not causally affect the SSW estimates and only indirectly ID through the density shapes of duty cycle durations.
[73] H. C. Gils, *Abschätzung Des Möglichen Lastmanagementeinsatzes In Europa. In 8. Internationale Energiewirtschaftstagung*, 8th Int. Conference on Energy Economics, Vienna, IEWT **2013**, http://elib.dlr.de/83717/.
[74] K. Appelhans, S. Exner, R. Bracke, *Analyse Des Deutschen Wärmepumpenmarktes – Bestandsaufnahme Und Trends. Technischer Bericht*, Internationales Geothermie Zentrum Bochum, University of Applied Sciences, Tech. Rep., 02 **2014**.
[75] Substituting into Equation (8,13,14,15) from the study by Ziras et al.,[39] who apply TB bounds derived for PSC with lockout[37] to SSW control.
[76] We have put the result curves from this experiment series into the Supporting Information, together with simulation plots.
[77] P. Barooah, A. Buic, S. Meyn, *48th Hawaii International Conference on System Sciences*, IEEE, Piscataway, NJ **2015**, pp. 2700–2709.
[78] A. Coffman, *IEEE Trans. Power Systems* **2020**, *36*, 2428, https://doi.org/10.1109/TPWRS.2020.3033380.
[79] Independently, the signal may be partitioned spectrally (FFT) and different frequencies be assigned to the best-suited DR resource classes.[77,78]
[80] S. H. Tindemans, G. Strbac, in *IEEE Power and Energy Society General Meeting (PESGM)*, IEEE, Piscataway, NJ **2016**, pp. 1.
[81] ASHRAE, *2009 ASHRAE Handbook : Fundamentals*, American Society of Heating, Refrigerating and Air-Conditioning Engineers: Handbook Fundamentals. Inc., Atlanta, GA **2009.**
[82] M. Arendt, Ph.D. Thesis, Karlsruhe Institute of Technology, **2001**.
[83] R. Kemna, J. M. Acedo. Average EU Building Heat Load for HVAC equipment. Technical report, European Commission, **2014**.
[84] K. P. Schneider, J. C. Fuller, D. P. Chassin, *IEEE Trans. Power Syst.* **2011**, *26*, 2425.
[85] A. Arteconi, N. J. Hewitt, F. Polonara, *Appl. Therm. Eng.* **2013**, *51*, 155.
[86] D. Patteeuw, G. P. Henze, L. Helsen, *Appl. Energy* **2016**, *167*, 80.
[87] The state observer equation is written for linear switching hazard—superscripted$^{(L)}$—and, for the sake of brevity, in continuous-time; for the discrete-time equations all input is assumed constant during one time step (zero-order hold).
[88] To monitor the SoC, rather than using the thermal leeway, we may integrate the load deviations directly. This requires identifying the baseline load accurately and adapting it to the current ambient conditions and thermostat bounds.
[89] J. Zheng, G. Laparra, G. Zhu, M. Li, *IEEE Trans. Control Syst. Technol.* **2020**, *28*, 1915.