Energy & Buildings 349 (2025) 116452

Contents lists available at ScienceDirect

Energy & Buildings

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

e-values based continuous-time model selection for residential electricity demand forecasts

Fabian Backhaus [a,*], Karoline Brucke [a], Peter Ruckdeschel [b], Sunke Schlüters [a]

[a] DLR-Institute of Networked Energy Systems, Carl-von-Ossietzky-Str. 15, Oldenburg, 26129, Germany
[b] Carl-von-Ossietzky University Oldenburg, Ammerländer Heerstraße 114-118, Oldenburg, 26129, Germany

ARTICLE INFO

Keywords:
e-values
Model selection
Hypothesis testing
Competing forecasting models
Energy demand forecasting

ABSTRACT

With the growing number of forecasting techniques and the increasing significance of forecast-based operation, particularly in the rapidly evolving energy sector, selecting the most effective forecasting model has become a critical task. In this context, the superiority of a forecasting model over its alternatives will, in general, hold—if at all—only on average (over time or across scenarios), and model selection typically results in a single static decision. Instead, enabling real-time decision making in the energy and building context, we introduce the concept of *e*-values-based decisions, which has recently gained massive attention in the field of statistics. We obtain continuous-time, method-blind, data-dependent decision rules, which take and revise their decisions along with the incoming information of forecast errors. Nevertheless, they still provide statistical guarantees, including a fixed decision risk over the whole period of time. We extend the use of *e*-values for times where no procedure is significantly superior to its competitor by developing a simple persistence approach that dynamically combines input forecasts to generate new fused predictions. To demonstrate the performance of our method, we apply it to building electricity demand forecasts based on different artificial intelligence-based models. Our *e*-selection procedure enhances our forecast accuracy by 16.3 % compared to the deviation of a single forecast to an all-knowing forecaster. Additionally, it improves the reliability of the forecast in a dynamic environment, offering a valuable tool for real-time decision-making in the energy sector.

1. Introduction

Due to the rising share of weather-dependent renewable energy in most energy systems, the scheduling and dispatch problem is getting more complex and relies on forecasts for both demand and generation. Especially in building energy management, forecasts play a major role in scheduling and optimizing the decentralized resources like energy storage, heat pumps, electric vehicles, and other flexible consumers [1]. Since energy management is mostly based on the respective day-ahead electricity market, the 24h forecast horizon is highly relevant in the building context [2]. Prediction-based demand side management is able to harness available decentral flexibility potentials [3] and enable, e.g., load shifting, etc. [4]. The quality of predictions is of great importance for the operational optimization under uncertainty in the building sector, as Schmitz et. al. show exemplarily in Schmitz et al. [5], optimizing the operation of a district heat pump. This results in two main research areas: The development of high-quality forecasting algorithms in general, and subsequently choosing the best forecasting method for a specific application over all or online. Many different forecasting methodologies exist with varying advantages and disadvantages. Statistical approaches like standardized load profiles [6], ARIMA [7], or naive persistence approaches typically have small computational cost but are often outperformed by more sophisticated approaches like Support Vector Regression (SVR) [8] and Artificial Neural Networks (ANNs) such as Long Short Term Memory (LSTM) [9], which on the other hand risk to be computationally very expensive. A recent study showed the application of different Reservoir Computing (RC) techniques for energy demand forecasting in the building context with high forecast quality and small computational effort [10]. Also, functional models are used to predict, e.g., the electricity price [11] or electricity demands [12]. Besides computational cost and forecasting quality, the need for expert knowledge or required amounts of data are additional criteria for evaluating the methodologies. Especially in the building context, every building has unique characteristics for instance, due to individual behavioral patterns in the residential context and vastly differing appliances and energy demands in commercial and industrial buildings, depending on the business model. Additionally, there can be many time-dependent changes in demands, e.g., daily patterns (morning to afternoon to

* Corresponding author.
E-mail address: fabian.backhaus@dlr.de (F. Backhaus).

https://doi.org/10.1016/j.enbuild.2025.116452
Received 23 October 2024; Received in revised form 22 August 2025; Accepted 14 September 2025
Available online 28 September 2025

List of symbols

Symbol	Description
α	significance level
$\mathbb{E}_{\mathbb{P}}$	conditional expectation with respect to distribution $\mathbb{P}$
$\mathcal{F}_t$	σ-algebra given the information up to time t
(p_t), (q_t)	forecast sequences
$\mathcal{H}_0(p,q)$	null hypothesis: forecast (q_t) is not better than (p_t)
$E_\lambda(t)$, $E^*_\lambda(t)$	e-process for $\mathcal{H}_0(p,q)$ or $\mathcal{H}_0(q,p)$
$\mathfrak{p}_t$, $\mathfrak{p}^*_t$	p-process corresponding to the e-process
δ_t, $\widehat{\delta}_t$	(empirical) score differentials
Δ_t, $\widehat{\Delta}_t$	(empirical) average score differentials
$\widetilde{\delta}_t$, $\widetilde{\Delta}_t$	bounded empirical (average) score differentials
ω	number of past observations. Symbols indexed by ω use the last ω observations
V_t	variance process of the score differentials
$\psi(\lambda)$	ψ-function
$u_{\alpha/2}$	uniform boundary at significance level $\alpha/2$
C_α	$(1-\alpha)$ confidence interval
θ_t	prediction using the e-procedure
$w_{p,t}$, $w_{q,t}$	weights for forecasts (p_t) and (q_t)
$f(x)$	sigmoid function to bound score differentials

nighttime), weekly rhythms due to changes from weekends to working days, or seasonal changes affecting the time spent indoors and in the northern hemisphere, especially heating demands. The heat demand data shown in Schmitz et al. [5] shows this very well with a vanishing space heating demand during the summer time and high demands in the winter with pronounced daily patterns due to the night-time reduction of indoor temperatures. With sector coupling technologies like heat pumps, this significantly affects the electricity demands. Thus, there is no best "one fits all" forecasting model for the building sector in terms of energy or electricity demands. Instead, forecasting models with different characteristics are competing against each other, but the model selection procedure tends to be a complicated multi-criteria decision process which we want to focus on in this paper. Despite being highly non-trivial, the model selection in the energy sector is mostly carried out ex-post by choosing the model simply with the smallest forecasting error. Such ex-post selection strategies have been pursued, e.g.,—for a variety of different error measures—in the review paper [13], as well as for predicting heating demands [14]. However, energy management in real-world systems like buildings is usually performed as a process in real-time, e.g., using model predictive control as in Kwak et al. [15]. Real time in the building context mostly refers to continuous time intervals of 15 min, depending on the respective electricity markets. For further examples of real-time energy management problems, see also Guo et al. [16], who perform real-time energy management for plug-in electric vehicles (EV), and Quan et al. [17], where a fuel cell EV is considered. This emphasizes the need for continuous real-time model selection of competing forecast techniques. Additionally, it is important to be able to quantify and limit the risk of decision-making for one or the other forecasting model in real-world energy management. Otherwise, a false decision—especially with unknown decision risk—could have negative economic implications for the different stakeholders, like distribution grid operators or the building owners. In the following subsection, a brief overview of existing model selection methods for competing forecasters is outlined.

1.1. Model selection methodologies

Compared to research on new forecasting methodologies, which is extensive throughout different domains of application, research on model selection approaches for competing forecasts has not yet gained a similar level of attention. This is particularly apparent in the field of applied research, and—in the perception of many applied scientists—has been lacking clear-cut principled guidelines to some extent [18].

Model selection in general can be framed and categorized in several respects. On the top level, one may distinguish model selection problems, where a given forecaster is to be enhanced by taking in additional information, and selection problems, where a set of forecasters is given and one wants to select the best among them. The former class of problems needs insights into the (structure of the) forecasters to be improved. In the machine learning context, this is closely related to transfer learning, see Pan and Yang [19]. We do not pursue this in this paper and rather take the competing forecasters as given and produce a series of rankings among them. So, to some extent, one might term this type of model selection as model ranking. For these rankings, we only use the past forecasts and the corresponding real values to select the best among them in real time. This allows us to be completely model-agnostic as to the competing forecasters.

Note that such a ranking decision among given forecasters does not make a direct judgment on the actual performance of the forecasters – if forecaster A is better than forecaster B, this does not imply that A is a good forecaster. For the scope of this paper, we assume that the set of competing forecasters has been chosen carefully so that one may expect at least one of these forecasters to perform well in absolute terms, too.

Within this ranking problem, we can distinguish settings according to the number of competing forecasters. To better convey the main ideas, this paper is limited to the case of two competitors, but we indicate possible extensions in the outlook of this paper. In case of ties, we may differentiate between "crisp" decisions and decisions where a fusion or combination of forecasts is allowed. Crisp in this context means searching or identifying the best alternative. Fusing forecasts for improved performance, on the other hand, is closely related to the general concept of ensemble methods, e.g., Wu and Levinson [20], Kumar et al. [21]. Our strategies ii) and iii) defined in Section 2.6 below could in fact be seen as particular ensemble methods. However, the framework pursued in this paper enhances these combined procedures by probability guarantees on the errors, which is (usually) not in the scope/focus of ensemble methods. Finally, the last distinction considered here concerns the type of decisions taken: Do we head for one selection (or fusion) decision once for all time, or for a stream of decisions which is allowed to sequentially take into consideration incoming information on prior forecast errors of the procedures? The main contribution of this paper is to head for general sequential decision streams combined with simultaneous error control, which is outlined in detail in Section 1.2.

As already mentioned above, most publications that perform model selection in the energy context compare error measures ex-post on a given test data set, which can be considered a naive benchmark method. However, due to the limited data availability at time t in the operation, this approach is not really applicable for real-time model selection. To overcome this, the authors in Swanson and White [22] use out-of-sample forecast-based model selection criteria for real-time macroeconomic forecasting. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also often used for model selection as in Billah et al. [23]. Besides the AIC, the authors of Liu et al. [24] additionally use the Pearson and Spearman correlation index for model selection.

Given the stochastic nature of the observations underlying model selection through a thorough approach would aim to control the stochastic uncertainty by means of probabilistic guarantees, such as specifying a significance or confidence level α. In a model selection context, such a guarantee means to select the forecast model with smaller error based on past average performance of a given time period while controlling the probability of a false selection.

Traditionally, these guarantees relied on strong structural assumptions on the distribution of the observations, such as stationarity, ergodicity, typically along with assumptions controlling the decay of dependence over time, such as mixing conditions or weak dependence conditions, as laid out in great generality in e.g. Bertail et al. [25].

In energy forecasting, though, typical seasonal effects and intra-day patterns indicate that stationarity cannot hold for the original observations– obviously, in Northern/Central Europe, in winter more energy is need for heating than in summer, and in private houses, you have daily recurrent peaks in demand when most people get up on business days and less energy demand when children are at school and adults to large extent are at work. You have well-predictable patterns distinguishing between workdays and bank holidays [6]. All these patterns can occur at fixed lags (e.g., if work routinely starts at 8 a.m.) or with stochastic periodicity (e.g., covering some persistent weather phenomena). Preprocessing techniques like detrending and deseasonalization can mitigate some of these violations, but other assumptions, such as ergodicity or weak dependence, remain challenging and hard to rigorously test.

Moreover, we are not heading for only one model selection decision "once for all" to be taken retrospectively and to be based on a fixed set of observations, but rather for continuous model selection in the same rhythm as the incoming new observations, which can be seen as a sequential decision approach as brought up in Wald's seminal paper [26]. Such approaches inherently must consider information simultaneously along a path of observations, which seems to suggest the need for even stronger structural assumptions, thus narrowing the application scope. Nevertheless, in particular settings, such continuous-time decision procedures have found their way to practical applications, as in the well-known CUSUM-type control charts (e.g., Vivancos et al. [27, Figure 2]) of stochastic quality control. However, when applied to dynamic settings with time-dependent observations, as in energy forecasting, the structural assumptions mentioned above traditionally could not be dispensed with.

With the recent advent of many new powerful procedures based on so-called "e-values", see Ramdas et al. [28], Shafer et al. [29], the necessity of these structural assumptions could almost entirely be dropped. Methods based on e-values are now also covering arbitrarily non-stationary situations in time-dynamic settings, therefore enabling continuous or real-time[1] Model selection of competing forecasting models. Additionally, e-values provide strong probabilistic guarantees, which so far has been out of scope for machine learning tools. Methods based on e-values were brought to continuous-time model selection in Henzi and Ziegel [30], Choe and Ramdas [31]. Contrary to Wald's setting [26], however, these methods head for streams of local decisions, valid only for the one-step ahead forecast, i.e., one does not stop (as Wald's procedure) after a taken decision but continues collecting the incoming evidence and revises former decisions continuously. Without going through the proofs of the validity of the procedures in Henzi and Ziegel [30], Choe and Ramdas [31], this makes it plausible why, for this purpose, stationarity is not needed, while for decisions on superiority valid "for all time" one evidently will have to ensure that non-stationary behaviour can be extrapolated well to a yet unseen future.

Technically, though, such methods amount to controlling whether a certain evidence measure remains within or crosses certain (simultaneous) control bounds, much the same way as in the mentioned control charts of CUSUM-type.

Despite the obvious advantages of continuous-time model selection, e-values so far have not been applied in the context of building energy management to the best of our knowledge. In addition, so far, e-values-based model selection as in Henzi and Ziegel [30], Choe and Ramdas [31] has been limited to predicting binary outcomes. So, in this sense, extending the applicability of e-values-based CUSUM-like charts for forecast model selection to the prediction of continuous outcomes like energy demand should be welcome.

When it comes to error assessment, the approach laid out in this work easily allows for quite general loss functions, including mean squared error (MSE), mean average error (MAE), and many more. This flexibility as to the choice of the loss does not say that the actual choice of the loss does not matter—to the contrary: different losses in general will lead to different rankings of forecasts. Therefore, it is crucial to use a loss function reflecting the needs of the decision makers in the respective domain, which is particularly true in the energy and building context. In this paper, we use MAE for the sake of reference and for ease of comparison to prior work in the context of building energy management, see Coignard et al. [32]. This choice does not imply any statement as to the preferability of MAE compared to other losses.

1.2. Contribution of this study

In this work, we make the following contributions to the field:

- We apply *e-values* for dynamic model selection in the building energy context for the first time to the best of our knowledge. With that, we are able to continuously rank two competing forecasting models while still providing statistical guarantees for our decisions. Note that in Section 6, we outline possibilities on how to extend the methodology to be able to rank $k > 2$ competing forecasters. Doing so, we are able to choose between forecasters in an online setting based only on available information in real-time. This is of high importance for real-world building energy management systems and the enhancement of their performance, since individually best forecasters can be chosen for each building for each time step.
- We propose and discuss three different strategies on how to combine forecasts when the e-values do not favor one of the considered models with significance. Two of these strategies can be seen as ensemble approaches themselves while still providing error control and fixed decision risks over the whole period of time.
- In a case study, we apply the proposed methodology for continuous model selection using two forecasting models, namely LSTM and Next Generation Reservoir Computing (NG-RC), which were applied on a publicly available electricity demand dataset on residential buildings in a previous study [10].

We stress the fact that the proposed model evaluation methodology does not come with any assumptions on the forecast methodologies and hence covers any forecasting techniques, including ensemble models. With regards to the considered forecasting models, e-values are completely model-agnostic. Additionally, the approach is very flexible with respect to the error measure for model evaluation and only needs very few assumptions. Due to the proposed approach being highly adaptable in terms of forecast models and error measures, it is of high importance for the building energy sector, where each building shows unique characteristics and users or residents have unique preferences or behavioral patterns.

As indicated, the key benefit of the approach is the probabilistic error guarantees for a given significance level α. Of course, the warranty given by the e-values comes with a price, and we cannot exclude that other decision tools might in some scenarios show better forecast performance, but then, based on the same available information, they cannot offer a probabilistic guarantee in the form of a (permanent and simultaneous) level α control. Despite this price, in the sequel, we will show that our procedure can be tuned to be competitive. Providing statistical guarantees in the context of building energy management is highly relevant for enabling respective business models in real-world systems.

This paper is structured as follows. First, in Section 2, we begin by introducing the mathematical concept of e-values, along with relevant definitions and notations from existing literature. We then detail the

[1] Strictly speaking our procedure described below comes in two phases: A calibration phase for tuning the hyperparameters (Step 1. in Section 2.6), which can be done offline, and a second phase (Steps 2. and 3.) with fixed hyperparameters, which can be done real-time, or more specifically in the same rhythm as the incoming forecasts, as the computation of the e-value-based decisions is cheap in terms of computation time.

construction of a sequential hypothesis test, which also leads to the development of confidence sequences for a given significance level. Building on this foundation, we introduce the e-selection procedure, which extends the sequential test to generate new predictions by dynamically combining the compared forecasting methods. Afterwards, in Section 3, we present the electricity demand time series data used in our application, along with the forecasting techniques employed. We also discuss data transformation methods and provide benchmark scores for comparison. In Section 4 we present the results of our study, including the performance of the e-selection procedure and an analysis of computational runtime. Finally, in Section 5, Finally, we discuss the limitations of our proposed procedure and provide further insights into the guarantees it offers. We also offer a brief outlook on potential directions for future research in Section 6.

2. Methodology

In this section, we briefly introduce the mathematical theory of e-processes, which provides a framework for so-called *Safe Anytime-Valid Inference* (SAVI). SAVI refers to statistical inference (i.e., tests, confidence bands, decisions) along with probabilistic error control in a sequential (possibly even continuous-time) setting, often without relying on distributional assumptions of the observed entities (in our case The *"Anytime-Valid"* in SAVI alludes to the key feature that e-processes allow for *optional stopping*, which is essential for the task to choose between two candidate forecast procedures in a continuous time setting under probabilistic error control as in Henzi and Ziegel [30]. Here, *optional stopping* refers to the fact that we monitor an evidence measure continuously over time along with the incoming observations, and *optionally* stop and make a decision, at a random time that is not known when monitoring starts, when sufficient evidence has accumulated so that the probability for a false decision can be kept controlled. In the context of continuous-time model selection, this decision will determine which forecast is superior to the other.

We begin by formally introducing e-values in Section 2.1. Subsequently, we set up corresponding test statistics based on scores for the competing forecasts in Section 2.4. These test statistics are used to distinguish formal hypotheses introduced in Section 2.5, leading to an e-selection procedure in Section 2.6.

2.1. Fundamentals for e-values

An e-process (E_t) is a nonnegative stochastic process[2] whose expected value $\mathbb{E}_P[E_\tau]$ is upper bounded by one for any arbitrary stopping time[3] τ under a given null hypothesis $\mathcal{H}_0$ [28]. Formally, this is expressed as:

$$\mathbb{E}_P[E_\tau] \leq 1 \tag{1}$$

for all stopping times τ and $P \in \mathcal{H}_0$.

Apart from Eq. (1), no further assumptions on the distribution or on the (stochastic) dependence of (E_t) over time are made, which makes e-processes particularly appealing for our time-dependent, non-stationary sequences of electricity demands.

A realization of an e-process is called e-value. This concept originates from the term, betting score" by Shafer in Shafer [34].

2.2. Relation to p-values

While a p-value represents the probability, under the assumption the null hypothesis holds, to observe, in a new experiment, a value for the test statistic being at least as large as the one observed, an e-value measures the accumulation of evidence against this hypothesis, growing rapidly when the hypothesis is violated [28,30]. As expectations, e-values can simply be combined by averaging them, with the average remaining an e-value, which is an important advantage over p-values in a sequential setting, compare [35].

2.3. Relevance of e-values

The relevance of e-values for testing lends to Ville's inequality, which entails the following bound valid for any e-process (E_t)

$$P(E_\tau \geq 1/\alpha) \leq \alpha \tag{2}$$

for all stopping times τ and $P \in \mathcal{H}_0$.

By means of Eq. (2), any e-process (E_t) can be translated into a sequential hypothesis test controlling the (familywise[4]) Type I error for a given null hypothesis $\mathcal{H}_0$, compare [28,37]. More specifically, we can reject the null hypothesis $\mathcal{H}_0$ at a familywise significance level of α as soon as E_τ surpasses the value $1/\alpha$ from below. While providing anytime-valid inference [28,30], e-values-based methods typically result in lower statistical power compared to those designed for fixed sample sizes (often called pointwise), but of course the pointwise error guarantee then only is valid for one time point, and combining pointwise guarantees for several time points bears the risk of alpha error cumulation [38]. The connection to p-values is given by the fact that, by taking

$$\mathfrak{p}_\tau = \min(1, 1/E_\tau), \tag{3}$$

any e-value can be converted into a conservative p-value [35].

As indicated, these e- and p-values are used to make decisions about which forecaster to prefer. To do so, we formally introduce the competing forecasts, the scores assessing their accuracy, and a (sequence of) test statistics based on the differences of these scores. In this dynamic setting, we must also carefully specify the amount of information available for the test statistics at a given time instant t. Much of this layout closely follows [31], who consider continuous time superiority testing for binary predictions. In fact, one contribution of ours is to extend their setting to superiority decisions for unbounded scores such as the mean absolute error (MAE).

2.4. Definitions and formalizations

Following the formulations in Choe and Ramdas [31], we start with two procedures issuing forecasts, denoted as (p_t), (q_t), indexed by a time index $t \in \{1, \dots, T\} \subset \mathbb{N}$, along with the corresponding outcomes (y_t) generated from the real distributions (r_t). We make no assumptions on the sample rate or require equidistant observations over time. To simplify notation, we drop the interpretation of t as a unit of time. Instead, t serves as a counting index for the considered predictions, and the time span between $t-1$ and t is not used in the analysis.

2.4.1. Information sets

To specify the available information at time t for a specific process, we use the filtrations defined in Choe and Ramdas [31]. A filtration $\mathcal{F}_t$ formally represents an *information set* containing all observable events or random variables available up to and including time t. For instance, $\mathcal{F}_t$ represents the *oracle* filtration, which reflects complete knowledge of all relevant information up to time t.

[2] For our purposes, a stochastic process is a stream of possibly time-dependent random variables indexed by their observation time t, and t may range in an ordered set of time points $\mathcal{T}$, $\mathcal{T}$ discrete or continuous. Klenke [33]

[3] A stopping time is a random time τ, for which for each time t, the accumulated information $\mathcal{F}_t$ at time t is sufficient to decide whether $\tau \leq t$. Klenke [33]

[4] The *familywise error rate* (FWER) is a key concept in multiple testing / simultaneous inference, compare [36]. It denotes the probability, under $\mathcal{H}_0$, to falsely reject at least one of the multiple hypotheses; this corresponds to simultaneous confidence intervals giving a probabilistic guarantee that the intervals *simultaneously* cover the unknown parameter or outcome with a given level. FWER control implies the weaker form of *false discovery rate* (FDR) control, which only warrants that the ratio of falsely rejected hypotheses stays below a given bound.

2.4.2. Error assessment

Our approach allows for quite general loss functions to assess forecast errors. This includes absolute error, relative error, squared error, or even asymmetric weights for positive and negative errors. In fact, the only requirement to enable a fair, method-blind comparison and ranking of forecasts is that the loss function be *quasi-convex* in the sense of De Finetti [39]. It implies that the further a competitor is from the optimum, the larger the loss. Ultimately, though, for our simultaneous confidence bounds below, we do need boundedness of the loss function, but wrapping a possibly unbounded quasi-convex loss (like, e.g., absolute error) by a suitable, bounded, strictly isotone function, such as unboundedness can be mitigated, see page 10 below.

As noted above, this flexibility does not imply that the resulting ranking of forecasts is independent of the chosen loss function. Rather, different losses in general will lead to different rankings; so it is crucial to use a loss function reflecting the needs of the decision makers.

As indicated in this paper, we use the MAE, but we are well aware that in other contexts, other loss functions will better reflect domain-specific needs. Examples might cover monetary values associated with forecast errors.

2.4.3. Scoring functions / scoring rules

In our setting, where we consider point forecasts, the loss function is called *scoring function* in the terminology of Gneiting [40], and is a function $S(x, y)$ where x is the value of the forecast and y is the actual outcome. More specifically, in the case of the MAE, $S(x, y) = |x - y|$. If instead of point forecasts, one heads for probabilistic forecasts, the loss function is called *scoring rule* in the terminology of Gneiting and Raftery [41], and argument x is the distribution-valued probabilistic forecast. In the case of a binary output, a probabilistic forecast can be summarized by a real-valued success probability. Consequently, the distinction between a scoring rule and a scoring function becomes minor and can be neglected, as was done in Choe and Ramdas [31]. Still, one should note that in the binary-outcome setting of Choe and Ramdas [31], $S(x, y) = |x - y|$ would not be *proper*, meaning that the perfect probabilistic forecast given by the true success probability can be beaten in this rule. However, this does not negatively impact our application, which deals with continuous outcomes.

The evidence for a specific forecast is evaluated through the empirical score differences, defined as

$$\hat{\delta}_t := S(p_t, y_t) - S(q_t, y_t), \tag{4}$$

which comes with a forecast counterpart

$$\delta_t := \mathbb{E}[S(p_t, y_t) - S(q_t, y_t) | \mathcal{F}_{t-1}],$$

taking into account the conditional expectation given the information up to time $t - 1$. Following Choe and Ramdas [31], instead of focusing on the varying sequence $\hat{\delta}_t$, we want to base our decisions on the average empirical score differentials

$$\hat{\Delta}_t := \frac{1}{t} \sum_{i=1}^{t} \hat{\delta}_i \tag{5}$$

together with the unobservable sequence of the average expected score differentials $\Delta_t = \frac{1}{t} \sum_{i=1}^{t} \delta_i$. The sequence Δ_t indicates whether one forecast outperforms the other one on average. Since the MAE is a negatively oriented scoring rule, negative score differentials indicate a preference for forecast (p_t).

Now in the approach in Choe and Ramdas [31] and also in most techniques discussed in Howard et al. [42], for powerful tests and hence efficient model selection, it is crucial to be able to use some exponential concentration bounds constructed in a similar manner as the Hoeffding or the Bernstein inequality. To this end, we require that the empirical score differentials $\hat{\delta}_t$ should be bounded, meaning we assume

$$|\hat{\delta}_t| \le \frac{1}{2} \text{ for all } t \ge 1.^{5} \tag{6}$$

Eq. (6) in general is violated for the MAE. To address this issue, we obtain bounded scores by transforming our empirical score differentials $\hat{\delta}$ using an appropriate sigmoid function f in Section 3.4.

2.4.4. Test statistics

To construct a suitable e-process, we follow the approach of Choe and Ramdas [31] based on Howard et al. [43], which involves using a cumulative sum process M_t, whose deviations from 0 we want to control over time. In our case $M_t = \sum_{i=1}^{t} \hat{\delta}_i - \delta_i$, respectively, or after transformation, $M_t = \sum_{i=1}^{t} \tilde{\delta}_i - \mathbb{E}[\tilde{\delta}_i \mid \mathcal{F}_{i-1}]$. From now on, we only work with the transformed score differences and drop the notational distinction between the transformed and untransformed score differences, and for better reference, we rather use the notation for the untransformed ones taken from Choe and Ramdas [31], Howard et al. [43].

A natural way to generate an e-process is through the exponential transform[6]

$$E_\lambda(t) = \exp\left(\lambda M_t - \psi(\lambda) V_t\right), \tag{7}$$

with appropriate choices for V_t, λ, $\psi(\lambda)$. In this formulation

- V_t is the variance process for M_t and can be interpreted as a measure of intrinsic time to quantify the deviations of M_t from zero. In our setting, we adopt the default variance process defined by Choe and Ramdas [31], which is given by:

$$\hat{V}_t = \sum_{i=1}^{t} (\hat{\delta}_i - \hat{\Delta}_{i-1})^2, \tag{8}$$

  where $\hat{\delta}_i$ and $\hat{\Delta}_{i-1}$ are defined in Eqs. (4) and (5). An illustrative example of the process $\hat{V}_t$ is shown in Fig. 1 (left).
- $\lambda > 0$ is a hyperparameter that controls the growth rate of the e-process $E_\lambda(t)$. It adjusts the impact of the deviations M_t by assigning them a weight. Specifically, larger values of λ increase the weights of extreme deviations. Therefore, λ also controls the willingness to take risk, compare [42,43]. Note that the procedure returns valid probabilistic guarantees for any $\lambda > 0$, which allows the user to adjust the procedure to domain-specific needs.
- $\psi(\lambda)$ is a cumulant-generating like function (CGF-like), that determines the rate at which the process M_t can grow in relation to the intrinsic time V_t. In Howard et al. [43], several useful ψ-functions are discussed, but we focus on the sub-exponential function given by:

$$\psi_E(\lambda) = -\log \qquad (1 - \lambda) - \lambda \tag{9}$$

  for all $\lambda \in [0, 1)$.

  In Fig. 1 (right) are shown the sub-exponential ψ-function ψ_E along with the sub-gaussian function ψ_N which is described in Choe and Ramdas [31], Howard et al. [43].

2.5. Sequential tests using e-Processes

Our goal is to sequentially test if one forecast outperforms the other on average, with a given significance level α. This means we want to test whether the sequence Δ_t is positive or negative for all timesteps t. Formally, the null hypothesis is defined as

$$\mathcal{H}_0(p, q) : \ \Delta_t \le 0 \ \forall \text{ times } t \ge 1. \tag{10}$$

[5] In the original reference, Eq. (6) is spelt out as $|\hat{\delta}_t| \le c/2$for some$c \in (0, \infty)$, but we use $c = 1$ and adapt the equations accordingly.

[6] Mathematically, if M_t forms a martingale, the exponential transform $e^{\lambda M_t}$ results in a submartingale due to Jensen's inequality. This submartingale can be transformed into a supermartingale $e^{\lambda M_t - \psi(\lambda) V_t}$ for appropriate choices of ψ and V_t. By the definition of this supermartingale [33], $E_\lambda(t)$ forms an e-process [43].

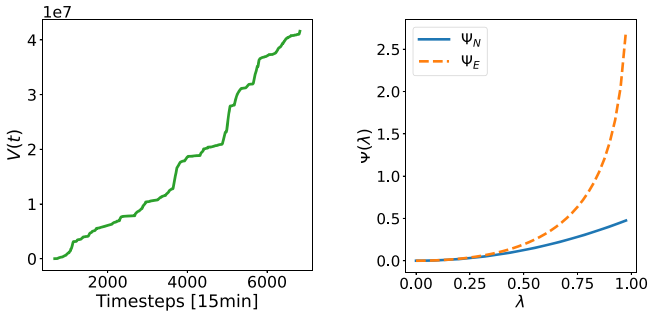

Fig. 1. Visualization of the required functions from Eq. (7). **(Left)** Illustration of the variance process $\widehat{V}_t$ using the data described in Section 3. **(Right)** Sub-exponential and sub-gaussian ψ-functions ψ_E and ψ_N for different $\lambda \in [0, 1)$.

For a negatively oriented scoring rule like the MAE, the hypothesis implies that the forecast (q_t) is not better than the forecast (p_t) on average across all times t. Analogously, by switching the order of the forecasts, we get

$$\mathcal{H}_0(q, p) : \ \Delta_t \geq 0 \ \forall \text{ times } t \geq 1. \tag{11}$$

The complete sequential test is then constructed by combining the two separate tests defined in (10) and (11), each with a significance level of $\alpha/2$. In our approach, we utilize Theorem 3 from Choe and Ramdas [31] as the main result. Accordingly the e-process for testing $\mathcal{H}_0(p, q)$ is defined as

$$E_\lambda(t) := \exp\left(\lambda t \widehat{\Delta}_t - \psi_E(\lambda)\widehat{V}_t\right) \tag{12}$$

for $\lambda \in [0, 1)$.

We write $E^*_\lambda(t)$ as the e-process for testing $\mathcal{H}_0(q, p)$. **Therefore the hypotheses $\mathcal{H}_0(p, q)$ or $\mathcal{H}_0(q, p)$ are rejected, if the corresponding e-process passes the threshold of $2/\alpha$ (Eq. (2)) from below.**

This construction of the e-process $E_\lambda(t)$ comes with time-uniform confidence sequences for Δ_t, which provide coverage guarantees that hold uniformly over time, ensuring that the process remains within the confidence interval at all times. This allows us to reject the hypotheses if the entire confidence sequence lies completely above or below zero [31]. Rewriting Eq. (2) and using the e-process from Eq. (12), we obtain the uniform boundary

$$u_{\alpha/2} = \frac{\psi(\lambda)\widehat{V}_t - \log(\alpha/2)}{\lambda}, \tag{13}$$

which then leads to the symmetric $(1-\alpha)$-confidence sequence for Δ_t:

$$C_\alpha(t) = \left[\widehat{\Delta}_t - \frac{u_{\alpha/2}}{t}; \widehat{\Delta}_t + \frac{u_{\alpha/2}}{t}\right], \tag{14}$$

such that

$$\mathbb{P}(\Delta_t \in C_\alpha(t)) \geq 1 - \alpha \ \forall t \geq 1. \tag{15}$$

This formulation ensures that the process Δ_t remains within the confidence sequence with a probability of at least $1-\alpha$ simultaneously for all time steps t.

A plot of $\widehat{\delta}_t$ (or $\widehat{\Delta}_t$) along with the confidence sequence $C_\alpha(t)$ over time can be seen as a variant of a CUSUM-type control chart, allowing for similar interpretation. Specifically, crossings of the $C_\alpha(t)$ bounds indicate critical events. For easier interpretation, we backtransform the score differences and differentials, as well as the upper and lower confidence bounds to the original MAE scale, which in our application is measured in Watts. An example of such a chart is shown in Fig. 7 (left).

2.6. e-selection procedure

In this section, we outline the procedure for selecting the forecasting method for future time steps using sequential tests, which we refer to as "e-selection". The goal is to construct a new prediction at each time step by combining two existing forecasting methods. The procedure can be roughly divided into three steps, which are performed at every time step t. The starting point is the empirical score differentials $\widehat{\delta}_t$ from Eq. (4), along with a predefined significance level $\alpha \in (0, 1)$. The procedure involves calculating the e-processes, conducting the corresponding sequential tests as a descriptive task, and combining the forecasting methods to generate a new prediction. The steps are summarized as follows:

1. *Calculate e-Processes:*
   First, we need to select the hyperparameter λ and the additional parameter ω, which represents the number of past observations to consider. Instead of accounting for the entire time horizon of the forecasting methods, we focus only on the most recent ω observations to better capture the dynamic behavior. Therefore the expression $\widehat{\Delta}_t$ in Eq. (5) is changed to rolling average

   $$\widehat{\Delta}_{t,\omega} = \frac{1}{\omega}\sum_{i=t-\omega}^{t} \widehat{\delta}_i \text{ for } t \geq \omega, \tag{16}$$

   and analogously $\widehat{V}_{t,\omega}$ and $E_{\lambda,\omega}(t)$ with the e-process started at time step $t-\omega$. By allowing optional stopping and continuation, the e-processes remain valid under this approach [30]. ω - or a grid of ω-values on which to evaluate the procedure - should be chosen ex ante according to domain-specific needs, for (each) given (value of) ω. Therefore, we recommend an offline hyperparameter optimization on a suitable representative dataset for λ. For a specific dataset, this will be discussed in Section 4. To obtain bounded score differentials $\widetilde{\delta}_t$, we use the transformation described in Section 3.4. We then use the values $E_{\lambda,\omega}$ and $E^*_{\lambda,\omega}$ at time step t for the sequential test.
2. *Sequential Test:*
   Having fixed ω and λ, we can take our sequential decisions in real-time i.e. for the test, we simply check whether the values $E_{\lambda,\omega}$ and $E^*_{\lambda,\omega}$ exceed the threshold of $2/\alpha$. We reject $\mathcal{H}_0(p, q)$, if $E_{\lambda,\omega} \geq 2/\alpha$, which indicates that forecast (q_t) outperforms forecast (p_t) on average over this specific period. Similarly, we reject $\mathcal{H}_0(q, p)$, if $E^*_{\lambda,\omega} \geq 2/\alpha$. In cases where neither hypothesis can be rejected, there is insufficient evidence to favor one forecast over the other, and we will discuss the subsequent steps in the following section.
3. *Generate Predictions by Combining Forecasting Methods (Forecast Fusion):*
   Again real-time, based on the results of the sequential test, we want to generate a combined prediction (θ_t), termed forecast fusion [44]. While the test itself does not rely on distributional assumptions, we do need some assumptions regarding future performance [31]. A simple approach is to use a persistence model, where if forecast (p_t) outperformed (q_t) on the previous day, we will also select forecast (p_t) for the next day. This ensures that (θ_t) is selected based on information available up to time $t-96$, assuming that the forecasts (p_t) and (q_t) are also $\mathcal{F}_{t-96}$-measurable. However, there are time steps where neither hypothesis can be rejected. For these cases, a different selection approach is necessary. Formally, the prediction at time step t can be expressed as:

   $$\theta_t = \begin{cases} p_t & \text{if } E^*_{\lambda,\omega}(t-96) \geq 2/\alpha \\ q_t & \text{if } E_{\lambda,\omega}(t-96) \geq 2/\alpha \\ z_t & \text{else,} \end{cases} \tag{17}$$

   where z_t is determined using one of the following three approaches:
   i) *Persistence:* The simplest option is the persistence e-selection. Once a decision is made at a time step $s < t$, an appropriate strategy is to continue using that selected forecasting method as long as there is insufficient evidence to switch. In this case,

   $$z_t = \theta_{t-1}. \tag{18}$$

   The advantage of this method is that it minimizes the number of switches between the possible predictions (p_t) and (q_t), leading to a more stable forecasting process.

ii) *Sampling:* The second strategy involves randomly sampling the forecast methods based on the amount of evidence, as indicated by the corresponding e-value. In this approach, we calculate the weight that determines the probability of selecting each method. Therefore, we use Eq. (3) to transform the e-values into anytime-valid p-values. We write:

$$\mathfrak{p}_t = \min(1, 1/E_{\lambda,\omega}(t))$$
$$\mathfrak{p}_t^* = \min(1, 1/E_{\lambda,\omega}^*(t)),$$

which results in the sampling weight for the forecast (p_t) as:

$$w_{p,t} = \frac{1 + \mathfrak{p}_t - \mathfrak{p}_t^*}{2} \tag{19}$$

and similarly, the weight for the forecast (q_t) is:

$$w_{q,t} = \frac{1 + \mathfrak{p}_t^* - \mathfrak{p}_t}{2} = 1 - w_{p,t}. \tag{20}$$

In this case

$$z_t = \begin{cases} p_t \text{ with probability } w_{p,t} \\ q_t \text{ with probability } w_{q,t}. \end{cases} \tag{21}$$

When $E_{\lambda,\omega}(t)$ and $E_{\lambda,\omega}^*(t)$ are both less than 1, indicating that there is insufficient evidence to reject either hypothesis, the weights reduce to 1/2 for each forecast. This results in a 50/50 chance of selecting either forecast p_t or q_t.

iii) *Weighted average (wAvg):* The weighted average approach combines the forecasts (p_t) and (q_t) based on the weight $w_{p,t}$ and $w_{q,t}$. Therefore

$$z_t = w_{p,t} \cdot p_t + w_{q,t} \cdot q_t. \tag{22}$$

As is well-known and can be easily proven, the sampling strategy ii) of selecting randomly among the forecasts is, in expectation, equivalent to the weighted average approach iii). However, the variance of approach iii) is strictly lower than that of approach ii), unless either $w_{p,t} \in \{0, 1\}$ or $p_t = q_t$ with probability 1. Therefore, when combining two forecasts at the same time t, the weighted approach generally provides increased statistical power. Still, approach ii) will be advisable, if constraints such as budget limitations preclude approach iii), limiting the number of forecasters to be used per time to just one.

To illustrate the steps involved in our procedure, we provide a visual overview in the form of a flowchart in Fig. 2. The diagram summarizes the e-selection procedure (Box E) as introduced in Section 2.6, highlights the initial choice of hyperparameters λ and ω (Box H), and indicates that the selection can be performed online in real time due to the very low computational cost per time step.

3. Data description and considered forecasting techniques

Having outlined our procedure for generating predictions by combining two forecasting methods, we proceed to demonstrate its practical application.

Specifically, we apply our procedure to forecast electricity demand using time series data. In this section, we introduce the electricity demand time series data and the forecasting techniques employed, which together form the basis of our application case.

While there are many promising forecasting models and algorithms in the literature, the focus of this paper is not on the forecasting method itself but on the process of dynamic real-time model *selection* among the given forecasts. The benefit of such a strategy becomes apparent in a situation where none of the competing forecasts is dominant in the sense that it is superior to its competitor(s) over the whole considered time period. This is the case for the study discussed in Brucke et al. [10], which was the starting point for this paper. In this paper, the authors compare and benchmark different recurrent network-based forecasting methodologies for household electricity demand. Their models result in forecasts with varying characteristics. More specifically, while the *Long-Short-Term-Memory* (LSTM) approach yields small MAEs with very smooth demand forecasts, the *Next Generation Reservoir Computing* (NG-RC) approach is able to follow more closely the wiggly behavior of energy demands with slightly higher MAEs on average. These different characteristics are reflected by changing superiority of the forecasts over time, and, consequently, neither the LSTM nor the NG-RC approach is dominant over the whole considered time period. So instead of a single static ex-post decision on superiority, we carry out the continuous model selection procedure based on e-values as laid out in the previous section with a fixed α guarantee level of 5 %. The remainder of this section is organized as follows: Section 3.1 shortly describes the raw data that was used to obtain the forecasts. The NG-RC and the LSTM approaches are briefly introduced in Sections 3.2 and 3.3, respectively. We summarize important statistical key features of all data sets and predictions in Table 1.

3.1. Raw data

The electricity demand time series used for forecasting is derived from the *EMS3* data set within the *EMSIG* data set [45]. This data set represents the energy data recorded by a decentralized household energy management system (EMS) from the DACH region, with a 15-minute resolution in Watts [W]. To take up our comment that data in building energy context as a rule are non-stationary, we fitted a generalized additive model (GAM) to our data using the `gam` function from the R package `mgcv`, Wood [46,47]. The model includes smooth terms for *time of day* and *calendar day*, and incorporates *weekday* as a categorical predictor with grouped levels. Applying the union bound, the probability of falsely rejecting at least one of these components does not have a linear effect that can be controlled at a level below 2×10^{-6}, thus amply rejecting stationarity.

Day-ahead predictions are created for the sum of the active power of electricity consumption, which is denoted by the column `sum_consumption_active_power` in the EMS3 data set. The models were trained and validated using the first 90 % of the data set, covering the period from January 1, 2019, to October 20, 2020, 20:15. The test set, which will be used for real-time forecasting comparisons, consists of the remaining 10 % of the data points, spanning from the evening of October 20, 2020, 20:30, to the evening of December 30, 2020, 22:45, and includes a total of $N = 6826$ measured electricity demand data points [10].

3.2. Next generation reservoir computing

Next Generation Reservoir Computing (NG-RC) is a machine learning algorithm that originates from nonlinear vector autoregression (NVAR), designed for analyzing dynamical chaotic systems using observed time-series information [48,49]. Unlike traditional reservoir computing, NG-RC constructs its feature vectors directly from unique polynomials of time-delayed input signals. This approach requires only a small amount of training data and yields computationally inexpensive optimization, which results in a highly efficient algorithm as is described in more detail in Brucke et al. [10].

3.3. Long-Short-Term-Memory neural networks

Long Short-Term Memory neural networks (LSTM), introduced by Hochreiter and Schmidhuber, are a special form of recurrent neural networks (RNNs) designed to handle the vanishing gradient problem of conventional RNNs. LSTMs fit very well for processing sequential data as they effectively capture long-term dependencies while retaining the ability to recognise short-term patterns [50]. This ability results from three categories of gates for each memory cell: Input, output, and forget gates. These gates regulate the storage and discarding of information

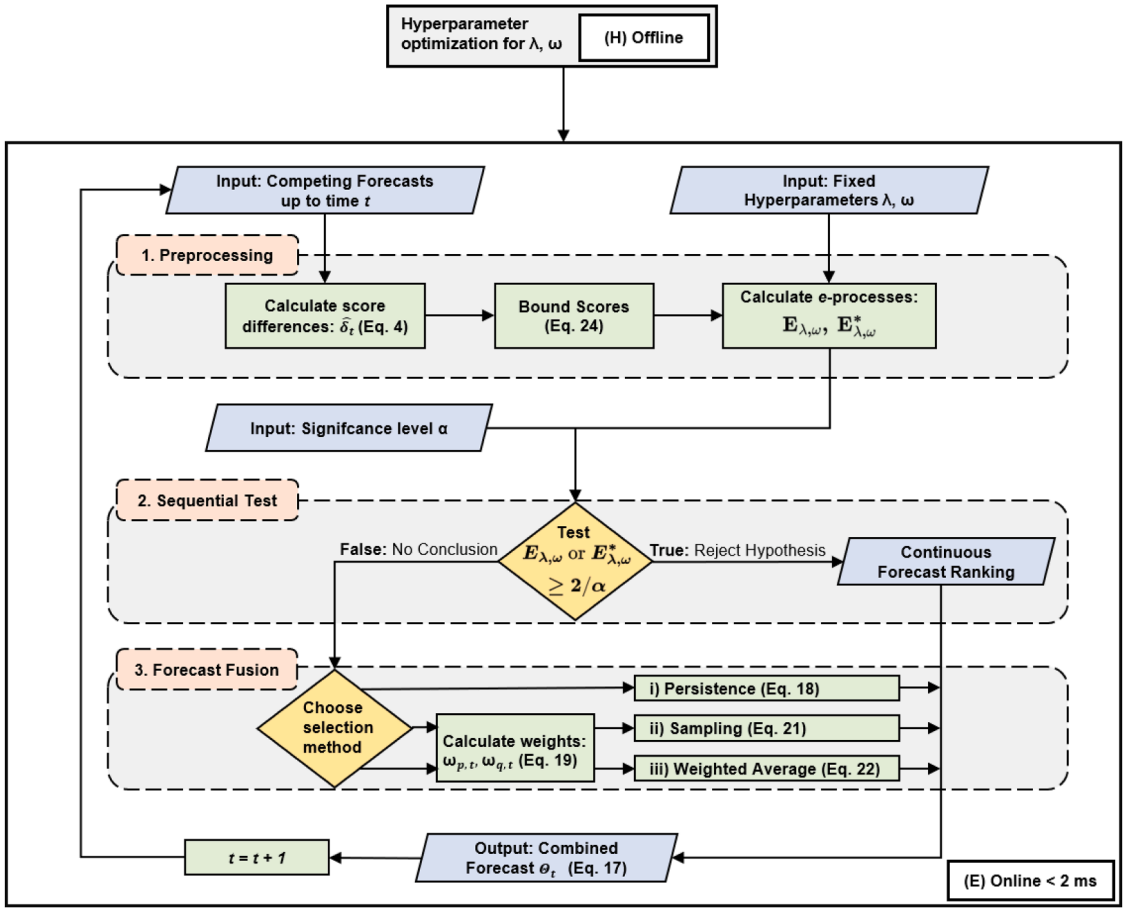

Fig. 2. Flowchart of the methodology for the e-selection procedure (Box E) from Section 2.6. While the forecast selection- and thus the computation of the corresponding e-values $E_{\lambda,\omega}$ and $E^*_{\lambda,\omega}$ is performed at each time step, the hyperparameters λ and ω are selected initially (Box H), as described in Section 4, and remain fixed throughout the entire selection procedure. The term *Online* in this context indicates that the procedure can be performed in real time as new data becomes available, owing to the small computational cost of less than 2 milliseconds for a single time step, as described in more detail in Section 4.

Table 1
Key statistical features of the measured load data set and the two forecasts based on NG-RC and LSTM. Oracle denotes the all-knowing forecaster selecting the best forecasting model at all times, which is the upper bound for the maximum forecast quality to be reached.

Statistics / Method	Actual Load	NG-RC	LSTM	Oracle
Data shape	6826×1	6826×96	6826×96	6826×96
Max. Load [W]	9346.00	6550.17	5001.01	/
Min. Load [W]	0.00	−2.06	165.02	/
Average Load [W]	757.91	725.73	546.35	/
Average MAE [W]	/	476.07	444.38	425.59
Max. absolute Error [W]	/	1116.99	1116.72	1116.72
Min. absolute Error [W]	/	177.08	129.26	129.26

and ensure that relevant data is retained over long sequences. Therefore, LSTMs are widely recognised as state-of-the-art for tasks such as time series prediction [10].

The actual electricity demand, along with the 96-step-ahead forecasts from the LSTM and NG-RC, is presented in Fig. 3 for the entire test set and two representative days. Additionally, Fig. 4 shows the corresponding histograms for the test set. Note that NG-RC predictions capture the real power distribution better compared to the LSTM predictions.

3.4. Data preprocessing

Each model generates a prediction for every time step of the test set for the following 24 hours, resulting in a prediction matrix $\hat{y} = (\hat{y}_{tk})$, where $t \in \{1, \dots, 6826\}$ denotes the time step and $k \in \{1, \dots, 96\}$ denotes k-step-ahead prediction horizon. The real power consumption is denoted by $y = (y_{tk})$ with the same dimension as $\hat{y}$. Every row in y contains the measured power consumption of the respective household from time step t for the next 24 hours. Accordingly, from one row t to the next row $t+1$, the data is shifted by one time step. We then use the MAE as a standard metric for evaluating load forecasts [51]. The MAE for each time step t is calculated using the following equation:

$$\mathrm{MAE}(\hat{y}_t, y_t) = \frac{1}{96} \sum_{k=1}^{96} |\hat{y}_{tk} - y_{tk}|, \tag{23}$$

where y_{tk} denoting the true realized value at time $t+k$. Let (p_{tk}) represent the prediction from NG-RC and (q_{tk}) the prediction from LSTM. We calculate the empirical score differences $\hat{\delta}_t$ between NGRC and LSTM as follows:

$$\hat{\delta}_t = \mathrm{MAE}(p_t, y_t) - \mathrm{MAE}(q_t, y_t).$$

Fig. 5 shows the histogram of the MAE scores for NG-RC (left) and LSTM (middle) along with the score differences $\hat{\delta}_t$ on the right-hand side of the figure. The overall mean value is shown by a dashed vertical line in every histogram. The score differences can be approximated by a shifted normal distribution.

To apply the procedure from Section 2.6, we need bounded scores like in Eq. (6), which can be obtained using an appropriate transformation function $f(x)$, which is chosen to be a sigmoid function in this work. More specifically, we require f to satisfy the following near-to-canonical conditions:

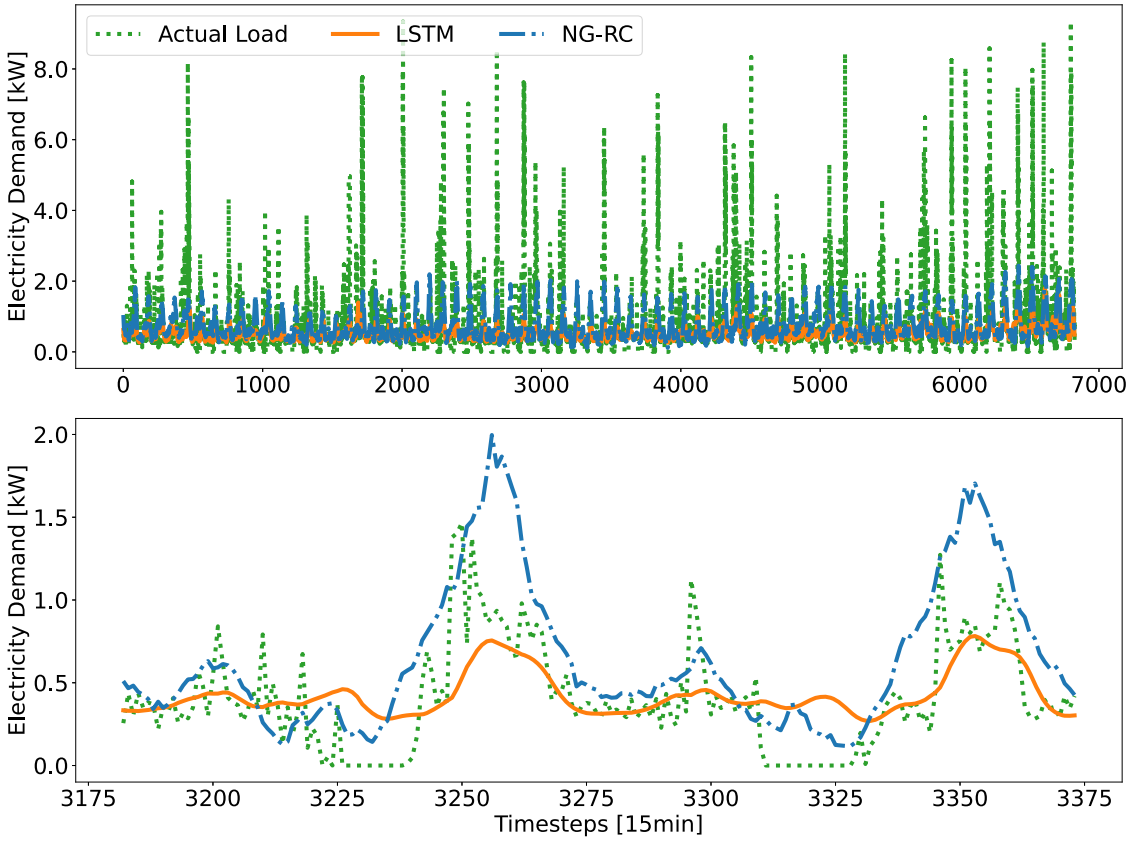

Fig. 3. Actual electricity demand (dotted line) and 96-step-ahead forecasts using LSTM (solid line) and NG-RC (dash-dotted line) in kilowatts. **(Top)** Full test set (6826 values) from October 20, 2020, 20:30 to December 30, 2020, 22:45. **(Bottom)** Detailed view of two days, from November 23, 2020, 00:00 to November 25, 2020, 00:00.

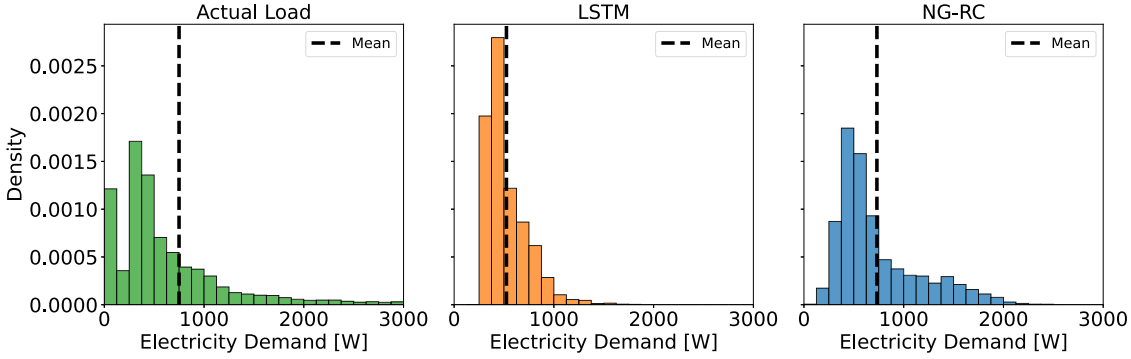

Fig. 4. Histograms of the actual electricity demand **(left)**, the 96-step-ahead LSTM forecasts **(middle)**, and NG-RC forecasts **(right)**, all in Watts. Dotted lines indicate the mean values. For improved visualization, the plots are truncated at 3000 W.

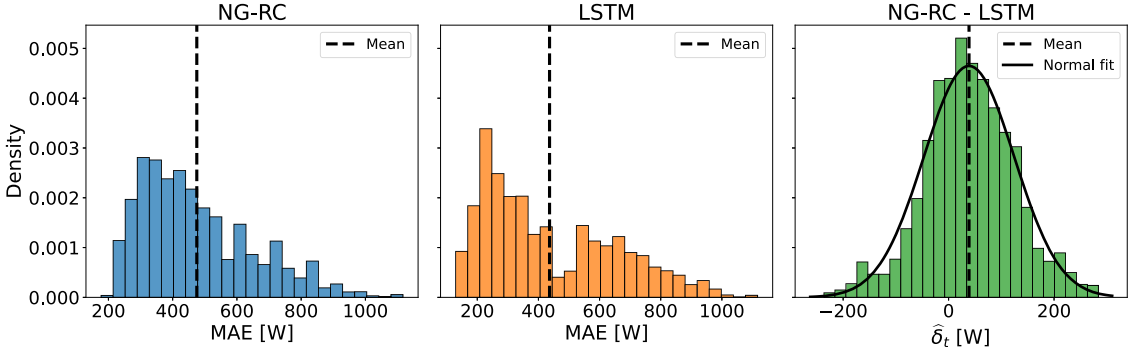

Fig. 5. Histograms of the MAE of NG-RC **(left)**, LSTM **(middle)**, and the empirical score differences $\hat{\delta}_t$ **(right)** with mean values (dotted line) in watts, and normal fit (solid line).

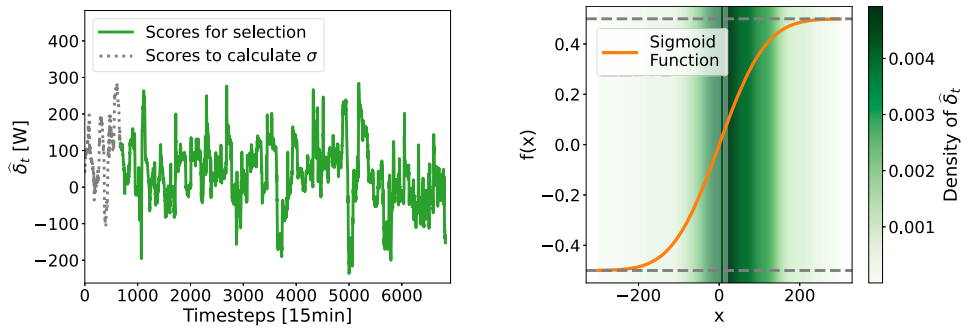

Fig. 6. (Left) Empirical score differences $\widehat{\delta}_t$ for the entire period. The dotted line represents the first week, which is used to calculate the scale σ. The remaining score differences are utilized for the selection procedure. **(Right)** Sigmoid function $f(x)$ in the range of the score differences $\widehat{\delta}_t$. The horizontal lines (dotted) represent the bounds at $-1/2$ and $1/2$. The color bar visualizes the quantity distribution density of the score differences $\widehat{\delta}_t$ along the x-axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(i) To minimize squeezing effects, f should be approximately linear in the central region where most of the probability mass is concentrated.
(ii) f should exhibit odd symmetry, with $f(0) = 0$.
(iii) To ensure boundedness, f should be curved in the tails.

The actual choice of f according to (i)–(iii) is of secondary importance, and different such choices will only lead to minor differences in the results. One possible such choice is

$$f(x) = \Phi(x/\sigma) - 1/2, \tag{24}$$

where $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution and σ is an appropriate scaling parameter.

The score differences $\widehat{\delta}_t$ for the entire data set are shown in Fig. 6 (left). σ is calculated as the standard deviation of the first week of $\widehat{\delta}_t$ (dotted line in Fig. 6 (left)). The remaining 6154 score differences are used for the selection procedure.

By definition, the transformed score differences $\widetilde{\delta}_t = f(\widehat{\delta}_t)$ fulfill the condition:

$$|\widetilde{\delta}_t| \leq \frac{1}{2} \ \forall \text{ timesteps t.} \tag{25}$$

Fig. 6 (right) represents the sigmoid function of Eq. (24). The dotted horizontal lines indicate the bounds at $-1/2$ and $1/2$. The color gradations visualize the quantity of score differences $\widehat{\delta}_t$ corresponding to the histogram on the right-hand side of Fig. 5. From this, we can visually verify that indeed f satisfies conditions (i)–(iii): Most of the score differences are centered around 0, with a slight rightward shift. This region is shaded dark green, indicating where the majority of score differences lie. Within this area, the sigmoid function is approximately linear as required. In contrast, outside this central region, represented by the lightly shaded green areas, the function exhibits strong curvature, approaching the limits of $\pm 1/2$ at both ends.

We aim at comparing the selection using e-values in Section 4 with the performance of the individual models NG-RC and LSTM. Additionally, we define an "oracle benchmark", which always selects the method with the lower MAE. Since the oracle represents a perfect selection, it establishes a lower bound for the best possible score achievable by our e-selection procedure. The average scores of the different benchmarks during the selection period (see Fig. 6 (left)) are presented in Table 1.

4. Results

In this section, we present the results of applying the e-selection procedure outlined in Section 2.6 to the electricity demand time series data described in Section 3. Specifically, we use the transformed score differentials $\widetilde{\delta}_t$ of the electricity demand forecasts and the e-process methodology to select a forecast model for every point in time. Doing so, we create a combined forecast which is benchmarked against each of the individual forecasting methods and against the oracle, which represents the best possible combined prediction. Our analysis focuses on the persistence e-selection method, including a hyperparameter optimization for λ and ω. Additionally, we report the computational time required for the selection processes.

Starting with the transformed score differentials $\widetilde{\delta}_t$, we first present exemplary results for a specific parameter combination of λ and ω. Specifically, we set $\lambda = 0.1$ and consider the data from the previous 7 days, corresponding to $\omega = 672$. Fig. 7 (left) displays the processes $\widetilde{\delta}_t$ (solid line) and $\widetilde{\Delta}_{t,\omega}$ (dotted line), along with the corresponding confidence intervals C_α for $\Delta_{t,\omega}$ calculated according to Eq. (14), at significance level of $\alpha = 0.05$ and $\lambda = 0.1$. The results are shown for two time periods of the data set: The days 7 to 14 of the data set are shown in the graphs at the top of Fig. 7 while the graphs at the bottom depict the days 62 to 69. The left y-axis in Fig. 7 (left) refers to the transformed scores, while the second y-axis reverts the scale of the bounded scores to the actual scores in Watts, indicating a linear transformation around zero, with higher absolute score deviations being compressed. The score differentials $\widetilde{\delta}_t$ exhibit significant variability, while the average score differentials $\widetilde{\Delta}_t$ seem to converge for each time period. The width of the confidence intervals decreases over time, eventually leading to the entire interval lying either above or below zero.

In Fig. 7 (right), the e-processes $E_{\lambda,\omega}(t)$ for the same time periods are plotted with the threshold value $2/\alpha$ on a logarithmic scale. In the first time period, the process exceeds the threshold at time step 852, indicating the rejection of the hypothesis $\mathcal{H}_0(p, q)$. This suggests that the forecasting method LSTM outperforms NG-RC during this time period. The same interpretation is supported by the confidence interval, which remains entirely above zero after this time. For the second time period, the e-process remains below the threshold and becomes very small. Therefore, we cannot reject the hypothesis $\mathcal{H}_0(p, q)$. To assess whether NG-RC outperforms LSTM, we should examine the process $E^*_{\lambda,\omega}(t)$ or consider the confidence interval, which remains entirely below zero after time step 6433. This indicates that NG-RC indeed outperforms LSTM during this period.

Applying the procedure to the entire dataset of 6154 prediction points results in an e-selection forecast θ_t for $t \in \{1, \ldots, 6154\}$. Fig. 8 illustrates the MAE of the NG-RC (dotted line) and LSTM (dashed dotted line) forecasts over the entire time period and depicts three different e-selection processes with varying ω and λ. The background area style in all three sub-figures indicates which model is selected by the e-process in that specific time period. Areas with diagonal lines represent the selection of NG-RC, while the dotted area indicates the selection of LSTM. Areas shaded with squares indicate time steps where no forecasting method is preferred, requiring the application of one of the approaches outlined in Section 2.6. Note that the first week is excluded from the predictions, as it is used to calculate the scale parameter σ for the transformation function $f(x)$ (Eq. (24)). In every plot, LSTM is the

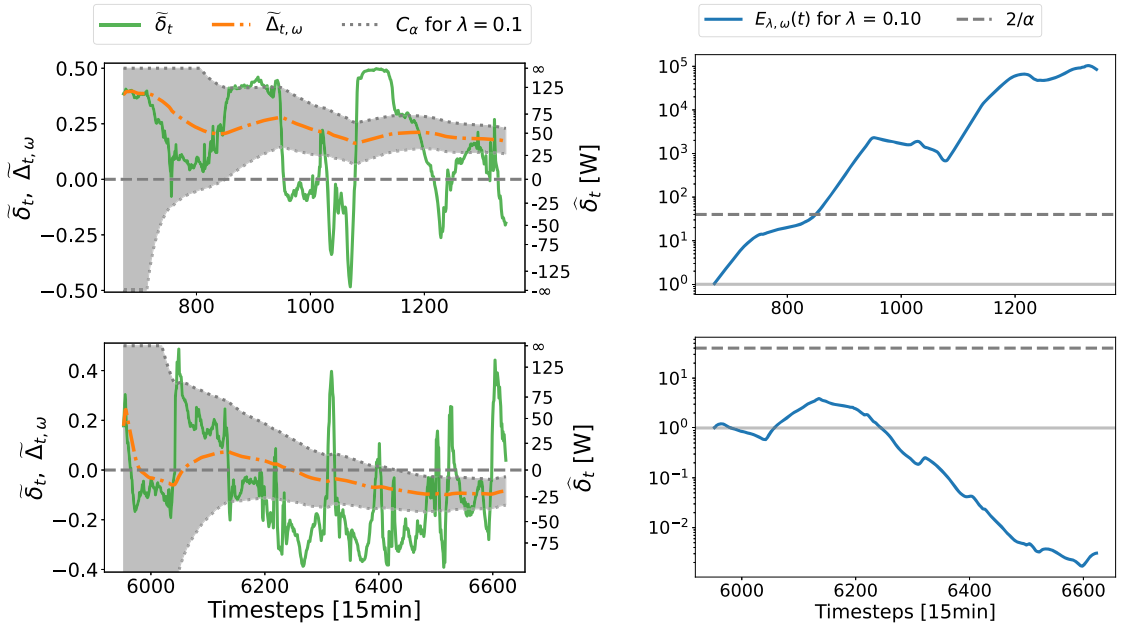

Fig. 7. Comparison of different processes for two time periods: days 7 to 14 (upper plot) and days 62 to 69 (lower plot). Confidence sequence and threshold at significance level $\alpha = 0.05$. **(Left)** Transformed score differentials $\widetilde{\delta}_t$ (solid line) and average transformed score differentials $\widetilde{\Delta}_{t,\omega}$ (dashed dotted line), along with the corresponding confidence sequence C_α (shaded area). The second y-axis represents the rescaled values from the first y-axis in Watts, showing a linear relationship around zero and a compression of higher deviations. The horizontal dashed line represents the value zero. **(Right)** e-process $E_{\lambda,\omega}(t)$ (solid line) with the threshold value $2/\alpha$ (dashed line) on a logarithmic scale. The horizontal solid line represents the starting value of one. Exceeding the threshold indicates that LSTM outperforms NG-RC during this period.

Table 2
Runtime of the different selection methods for various ω for the whole dataset.

	runtime e-selection [s]		
ω [days]	Persistence	Sampling	wAvg
1	2.46	6.36	2.46
7	7.70	7.79	7.80
14	12.50	12.72	13.12

most frequently selected model. However, NG-RC is primarily chosen towards the end of the time series for each plot. Notably, smaller rolling windows ω and higher risk tolerances λ result in more frequent and faster switches between models. Conversely, a larger rolling window of 14 days typically leads to only a single switch, occurring at the end of the time series.

To consider the computational cost, the runtime for each approach for the entire dataset is presented in Table 2 across various window sizes ω. All computations were performed on a Windows Server 2019 machine equipped with an Intel Xeon E5-2630v4 CPU and 256 GB DDR4-2400 RAM. The code was executed with Python 3.10, running on a single core without utilizing multiprocessing. Larger windows result in longer runtimes due to the increased computational demand of summing over more time steps. For instance, with $\omega = 7$, the e-selection procedure for a single time step requires less than 2 milliseconds.

Due to the relatively low computational costs of applying e-processes for model selection, we perform a hyperparameter optimization considering ω and λ. This optimization is carried out using both a simple grid search and the Python module `optuna` [52]. The parameter ranges considered are:

- $\omega \in \{1\text{h},\ 2\text{h},\ 1\text{d},\ 2\text{d}, \ldots, 14\text{d}\}$, where h denotes the hours and d denotes the days,
- $\lambda \in \{0.01, \ldots, 0.99\}$.

For each combination, the overall average score across the entire prediction period is calculated for each of the three e-selection methods. Combinations where no conclusion could be reached after the first time step of the first week were excluded from consideration to maintain consistency in comparison with the persistence method. Selected results for $\omega \in \{1\text{d},\ 4\text{d},\ 7\text{d},\ 14\text{d}\}$ and $\lambda \in \{0.1,\ 0.5,\ 0.9\}$ are presented in Table 3. For comparison, we benchmark these scores against the average scores of NG-RC and LSTM in Table 1. Scores that are lower than those of NG-RC and LSTM are highlighted in the table in bold font. The best achieved score was 441.31 W using the persistence method with $\omega = 7$ days and $\lambda = 0.07$. Although a deviation of 3.07 W from the LSTM model may seem minor, this actually represents a 16.3 % improvement compared to the LSTM's deviation from the best possible score of 425.59 W (oracle). With this parameter configuration, the e-selection method chooses the better forecasting model in 70.91 % of the time steps.

To examine all combinations, Fig. 9 shows a heatmap representing the deviation of the e-selection persistence method compared to the LSTM method across various parameter combinations. Combinations with window sizes of one and two hours were excluded from the heatmap because they couldn't select a method at the first time step. Cells marked with a, + " indicate an improvement of the combined forecast using the e-selection method. Choosing window sizes larger than five days consistently results in an improvement regardless of the choice of λ. The worst performance occurred with $\lambda = 0.09$ and $\omega = 1$d, showing a worsening of the forecast performance of -11.67 W. Overall, in 74.24 % of the cases, we get an improvement using the proposed e-selection persistence method. We conclude that ω needs to be larger than a minimal threshold – in our case 4 days in order to outperform LSTM. But the benefits of the hybrid procedure again become less pronounced for larger values of ω – in our case $\omega > 10$ days. At the same time, the choice of λ is largely insensitive, so –at least in our example– one could recur to other criteria, e.g. choosing λ such that in a pay-per-use regime with lower prices for NG-RC than for LSTM, one would use NG-RC more often.

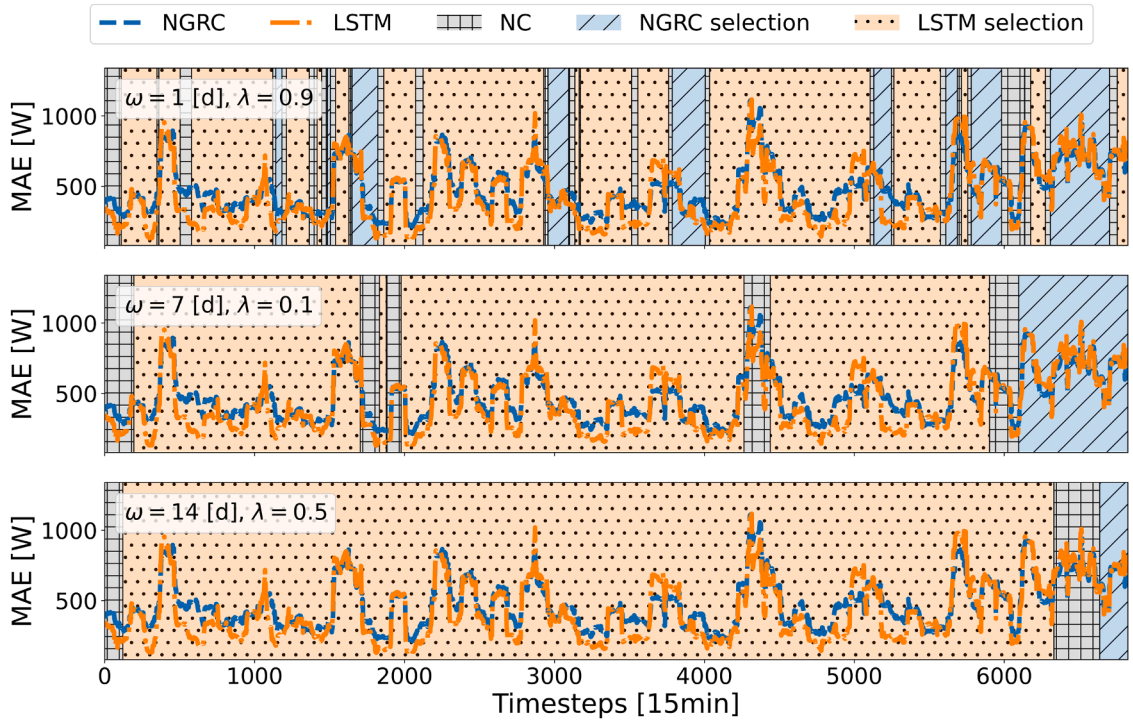

Fig. 8. MAEs for NG-RC (dashed line) and LSTM (dashed dotted line) over 6154 prediction points, excluding the first week of data. Shaded areas indicate model selection by the e-selection method for different hyperparameters ω and λ: diagonal lines for NG-RC, dotted for LSTM. Squares show periods where no clear preference is given.

Table 3
Average scores in Watts for the three e-selection approaches from Section 2.6 across various parameter combinations. Scores lower than those of NG-RC and LSTM in Table 1 are highlighted.

Hyperparameters		Average score [W]		
ω [days]	λ	*Persistence*	*Sampling*	*wAvg*
1	0.1	449.30	450.90	451.01
	0.5	449.35	448.96	449.17
	0.9	449.21	448.96	449.19
4	0.1	**441.81**	446.66	446.73
	0.5	445.56	446.50	446.54
	0.9	445.41	446.39	446.54
7	0.1	**441.48**	**442.12**	**442.26**
	0.5	**441.74**	**442.44**	**442.41**
	0.9	**441.67**	**443.01**	**442.96**
10	0.1	**442.83**	**442.44**	**442.47**
	0.5	**442.53**	**442.45**	**442.50**
	0.9	**442.64**	**442.48**	**442.51**
14	0.1	**443.96**	**443.27**	**443.30**
	0.5	**444.04**	**443.20**	**443.28**
	0.9	**443.97**	**443.03**	**443.20**

For the sampling method, we observe that 64.94 % of the cases result in an improvement, with the best performance achieving a score of 441.42 W and a deviation of 2.96 W. The worst performance for this method results in a score of 450.94 W, leading to a deviation of -6.56 W. For the weighted average method, 64.87 % of the cases show an improvement, with the best deviation at 2.91 W and the worst deviation at -6.92 W. Although these methods show a slightly lower percentage of improvements compared to the persistence method, they also show significantly better worst-case performances.

5. Discussion

Our results from the previous section suggest that the e-selection procedure, particularly the persistence approach in our application, can outperform a fixed choice of a forecasting technique (one time for all). Even in the worst-case scenario, it provides an improvement over consistently choosing the NG-RC forecast. While the parameters λ and ω control the frequency of model switching, for our data, the e-selection's performance seems to indicate a stronger dependence on the window length ω than on the value of the risk aversion parameter λ. Given that residential load profiles often exhibit daily and weekly patterns, it is plausible that the past $\omega = 7$ days represents the most critical time window for energy forecasting [53]. Notably, setting $\omega = 1$ step and $\lambda \approx 1$ closely approximates a true day-ahead persistence model. Rather than dividing the data into a separate validation set, our study demonstrates the application of e-values in a continuous time setting, which enables real-time model selection, including guarantees and known decision risk in the energy context. This is highly relevant to the building energy context since it enables individual online decisions only with the information available at that point in time. Individual here is two-folded. On the one hand, it refers to individual decisions for each time step in one building or a closed unit within a building. This is important since different forecasting models can have different performances throughout the day or seasons, as can be seen in Fig. 3. On the other hand, individual decisions can be taken with respect to different buildings reflecting the unique characteristics of energy demands, which could favor different forecasting models that are better at capturing the unique behaviors.

Although computational time is not the main focus of this work, it remains crucial for real-time applications. Even though the selection procedure itself is computationally efficient and can potentially be improved, generating two forecasts at each time step can be expensive, particularly when using computationally intensive methods like LSTM [10]. In the building context, where the time intervals for energy

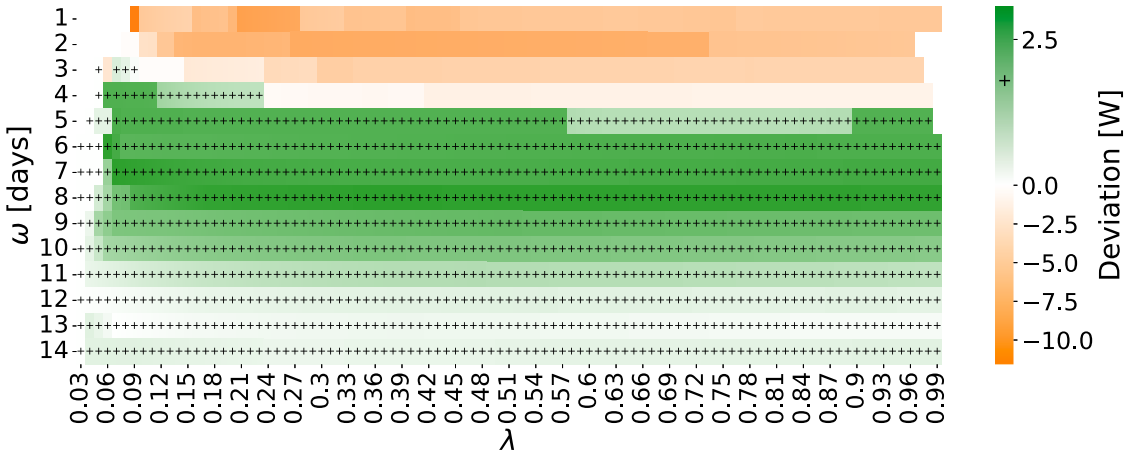

Fig. 9. Deviation of LSTM and the e-selection persistence method of the average scores across combinations of the hyperparameters ω and λ. Cells marked with a " + " indicates an improvement of the performance using e-selection.

management mostly are between 5 min and 15 min depending on the respective electricity market, the runtime of the model *selection* will not be a limiting factor as the runtime of one single e-selection process takes less than 2 milliseconds. The runtime for conducting the actual forecasts will be the main limiting factor for real-time applications in the building energy context. However, there are also smaller time intervals to be considered in the energy domain in general, e.g., when looking into frequency reserve.

Regarding the different e-selection approaches, our example does not provide conclusive evidence favoring one method over another. For $\omega \geq 7$, where our model consistently outperforms both LSTM and NG-RC, the average scores remain extremely close, varying by less than 4.7 % in comparison to the optimal score. Instead, one may focus on interpretational aspects. The persistence approach leads to longer periods using a single forecast, while the sampling and weighted average methods combine both predictions to enhance forecast accuracy. In this context, the weights $w_{p,t}$ can be interpreted as Bayes factors [54]. If a linear combination of the forecasts is allowed, the weighted average procedure is applicable. Otherwise, if a clear decision between p_t and q_t must be made at each step, the sampling approach is appropriate.

Despite LSTM outperforming NG-RC in the long term with respect to MAE on this dataset, NG-RC achieves a lower root mean squared error (RMSE), indicating that the selection outcome is highly dependent on the chosen scoring function [10]. This emphasizes the importance of carefully choosing a suitable error measure for each application, dataset, or building, respectively. On the other hand, this also shows the flexibility of the approach, enabling tailored solutions for the individual applications. We already discussed the possible effects and the importance of the chosen score in detail in our methodology in Section 2.4.2. A key advantage of using e-values is their ability to control the familywise error rate in sequentially dependent data, providing guarantees for the statistical decisions. In our study, we test $\widetilde{\Delta}_{t,\omega}$ at each time step t, offering a weaker guarantee than testing the score differentials δ_t across the entire dataset, hence is able to significantly reject the hypotheses more often (maintaining level-α-control). However, this guarantee applies strictly to the statistical test, not to the overall prediction, which requires additional assumptions. Consequently, it is challenging to verify whether the confidence level $1 - \alpha$ is actually achieved [31].

Limitations: While our approach is extremely general with no assumptions at all on stationarity, ergodicity, range of the forecast values (which is a clear breakthrough as to probability guarantees), one could consider it a limitation that the number of transitions/switches from one procedure to the other one does not enter the decision rules. In principle, if switching involves expensive conversion steps, the resulting procedures introduced in this paper could end up with too many such transitions and end up suboptimal as to costs. However, this is no knock-off drawback, as such conversion costs could be integrated as a penalty into the loss function. The modified procedure would only require transitions if the induced conversion costs are amortized by the benefits of the transition. Another limitation of our approach lies in the fact that we issue streams of superiority decisions, each of them only valid locally in time. On the one hand, this local validity adds flexibility when one is allowed to decide to use forecaster A in this period and forecaster B in the other period. But these local decisions will, in general, be of limited help for investment decisions, whether one should buy forecaster A or forecaster B. For such purposes, a global superiority criterion would be needed, so the local decisions would need to be aggregated in a suitable way; discussion of such aggregation mechanisms would be a topic for another paper. Finally, our setting so far only considers the selection between two competing forecasting techniques. A natural generalization would enlarge the set of competing techniques to $k \geq 2$ competitors. Remarkably, the structure of the decisions in this more general setting will remain the same as the ones discussed in this paper, but two adaptations will need to be made: (a) Refined simultaneous confidence bounds will be needed, which take into account that the best procedure must be better than $k - 1$ instead of 1 competitor[s], and (b) the off-diagonal terms in the respective confusion matrix gathering the occurrence probabilities of each false pairwise ordering in case of k competitors will have to be aggregated and weighted with respective costs in a suitable way. In fact, in the framework of classical sequential decision problems, this amounts to passing from the two-armed bandit problem (the analogue of which was the subject of this paper) to the k-armed bandit problem. The e-value approach has already come up with solutions for such problems, see Ramdas et al. [28, Section 7]. In particular, Kaufmann and Koolen [55] proposes solutions for (a) and (b) in the k-armed-bandit problem, albeit in the much narrower world of decisions for processes with distributions in an exponential family. Moreover, with additional notational and computational complexity, one could even allow for a time-varying number of competitors, i.e., k depending on t in a stochastic, "prequential" way in the sense of Dawid [56]. Still, the mathematics behind these generalizations would clearly go beyond the scope of this paper, and we instead refer to future papers.

From the building energy perspective, there are only very few limitations due to the high flexibility towards the forecasting models, the error measures, and the fast computational time of e-values. However, two points remain noteworthy. Firstly, we only choose the forecast model for the next time step and are not able to make statements on the superiority of different forecast models over all, thus on a global level or for greater time horizons. Instead, the proposed approach based on e-values provides a continuous stream of decisions for the very next time step

based on currently available information. For global decision-making, well-established ex-post model comparison methods can be used. Secondly, the proposed methodology only conducts relative comparisons of forecasting models and makes no statements on the absolute quality of the predictions themselves. If the e-values favor one model A over another B, that does not imply that A is automatically a good forecasting model. However, it can be assumed that the forecasting models that are taken into account for comparison are carefully chosen based on expert knowledge for the specific application or building. For absolute statements on the forecast quality, respective error measures have to be evaluated directly.

6. Conclusion and outlook

In this work, we successfully transfer and translate the e-value-based approach for time-continuous forecast model selection from Henzi and Ziegel [30], Choe and Ramdas [31] from binary outcome prediction to the energy domain. We apply the approach to residential building electricity demand forecasts. We extend this forecast model selection approach to forecast fusion/combination by specifying a combination of the two forecasts into a new, better one. Our study demonstrates that the e-selection method provides competitive results with minimal additional computational costs while offering a statistical guarantee. However, it's important to note that this guarantee does not imply that the sequential test will make the correct decision at every time step with an error rate of α, because this would involve information on the outcome distribution unknown to the decision process. Instead, it controls the average score differentials with a significance level of $1-\alpha$ [31]. The passage from the binary outcome setting of Henzi and Ziegel [30], Choe and Ramdas [31] to the continuous outcome setting of energy demand also requires the usage of scores adapted to this scale. To this end, we replace the Brier scores used in Henzi and Ziegel [30], Choe and Ramdas [31] with the MAE, which is a well-established error measure in the energy context. To achieve this, we suitably transform the MAE scores in an order-preserving manner. For interpretability, in our decision plots, we back-transform the results and confidence bounds to the original MAE scale. Even though we use the MAE exemplarily in this work, our proposed method on dynamic model selection works with many other error metrics as long as they are *quasi-convex*. Additionally, any forecasting techniques are possible for the model evaluation procedure since e-values are "method-blind". This makes the model selection approach based on e-values highly relevant and applicable to the energy context and especially the building domain, where unique characteristics apply in each building. On the real data example of a time series of residential electricity demand, these tests are shown to have enough power to obtain a clear ordering of the considered forecasters most of the time, which, in addition, is supported by a probabilistic guarantee.

Open Ends: As indicated, the probabilistic guarantees do not necessarily extend to a retrospective backtesting perspective, so future work should focus on properly backtesting the results presented in this study. The model-free and e-process-based backtesting procedure introduced in Wang et al. [57,58] offers a promising direction for such a validation. Additionally, further studies are needed to evaluate the performance and robustness of our e-selection method across different datasets and scoring rules. This includes validation of the hyperparameters involved. Moreover, the selection procedure should account for the varying computational costs associated with different forecasting methods. Passing from point forecasts to probabilistic forecasts, it is clear that one could easily proceed in the same way as in this work, simply replacing the MAE score function with the terms of a continuous ranking probability score (CRPS) as discussed in Gneiting and Katzfuss [59]. To this point, our proposed model selection framework based on e-values is limited to $k = 2$ forecasters but could be extended to $k > 2$ in future work as outlined in our discussion section. Currently, the number of transitions between forecasting models is not limited, but one could incorporate the associated "costs" of switching forecasting models easily into the penalty function. This could be covered by future work.

CRediT authorship contribution statement

Fabian Backhaus: Writing – original draft, Writing – review & editing, Formal analysis, Software, Methodology, Visualization, Validation, Data curation; **Karoline Brucke:** Writing – original draft, Supervision, Writing – review & editing, Resources; **Peter Ruckdeschel:** Writing – original draft, Writing – review & editing, Methodology, Conceptualization, Supervision; **Sunke Schlüters:** Supervision, Project administration, Funding acquisition.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

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

References

[1] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205.
[2] C. Deb, F. Zhang, J. Yang, S.E. Lee, K.W. Shah, A review on time series forecasting techniques for building energy consumption, Renew. Sustain. Energy Rev. 74 (2017) 902–924.
[3] H. Golmohamadi, Demand-side flexibility in power systems: a survey of residential, industrial, commercial, and agricultural sectors, Sustainability 14 (13) (2022) 7916.
[4] O.M. Babatunde, J.L. Munda, Y. Hamam, Power system flexibility: a review, Energy Rep. 6 (2020) 101–106.
[5] S. Schmitz, K. Brucke, P. Kasturi, E. Ansari, P. Klement, Forecast-based and data-driven reinforcement learning for residential heat pump operation, Appl. Energy 371 (2024) 123688. https://doi.org/10.1016/j.apenergy.2024.123688
[6] N. Pflugradt, P. Stenzel, L. Kotzur, D. Stolten, LoadProfileGenerator: an agent-based behavior simulation for generating residential load profiles, J. Open Source Softw. 7 (2022) 3574.
[7] L.C.M. de Andrade, I.N. da Silva, Very short-term load forecasting based on ARIMA model and intelligent systems, in: 2009 15th International Conference on Intelligent System Applications to Power Systems, IEEE, Curitiba, Brazil, 2009, pp. 1–6.
[8] Y.-C. Guo, D.-X. Niu, Y.-X. Chen, Support vector machine model in electricity load forecasting, in: 2006 International Conference on Machine Learning and Cybernetics, IEEE, Dalian, China, 2006, pp. 2892–2896.
[9] K. Brucke, S. Arens, J.-S. Telle, T. Steens, B. Hanke, K. von Maydell, C. Agert, A non-intrusive load monitoring approach for very short-term power predictions in commercial buildings, Appl. Energy 292 (2021) 116860.
[10] K. Brucke, S. Schmitz, D. Köglmayr, S. Baur, C. Räth, E. Ansari, P. Klement, Benchmarking reservoir computing for residential energy demand forecasting, Energy Build. (2024) 114236.
[11] F. Lisi, I. Shah, Joint component estimation for electricity price forecasting using functional models, Energies 17 (14) (2024). https://doi.org/10.3390/en17143461
[12] I. Shah, F. Jan, S. Ali, Functional data approach for short-term electricity demand forecasting, Math. Probl. Eng. 2022 (1) (2022) 6709779. https://doi.org/10.1155/2022/6709779
[13] Y. Sun, F. Haghighat, B.C.M. Fung, A review of the-state-of-the-art in data-driven approaches for building energy prediction, Energy Build. 221 (2020) 110022. https://doi.org/10.1016/j.enbuild.2020.110022
[14] Y. Guo, J. Wang, H. Chen, G. Li, J. Liu, C. Xu, R. Huang, Y. Huang, Machine learning-based thermal response time ahead energy demand prediction for building heating systems, Appl. Energy 221 (2018) 16–27. https://doi.org/10.1016/j.apenergy.2018.03.125
[15] Y. Kwak, J.-H. Huh, C. Jang, Development of a model predictive control framework through real-time building energy management system data, Appl. Energy 155 (2015) 1–13. https://doi.org/10.1016/j.apenergy.2015.05.096
[16] N. Guo, X. Zhang, Y. Zou, L. Guo, G. Du, Real-time predictive energy management of plug-in hybrid electric vehicles for coordination of fuel economy and battery degradation, Energy 214 (2021) 119070. https://doi.org/10.1016/j.energy.2020.119070
[17] S. Quan, Y.-X. Wang, X. Xiao, H. He, F. Sun, Real-time energy management for fuel cell electric vehicle using speed prediction-based model predictive control considering performance degradation, Appl. Energy 304 (2021) 117845. https://doi.org/10.1016/j.apenergy.2021.117845
[18] A.T. Tredennick, G. Hooker, S.P. Ellner, P.B. Adler, A practical guide to selecting models for exploration, inference, and prediction in ecology, Ecology 102 (6) (2021) e03336. https://doi.org/10.1002/ecy.3336

[19] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[20] H. Wu, D. Levinson, The ensemble approach to forecasting: a review and synthesis, Transport. Res. Part C: Emerg. Technolog. 132 (2021) 103357. https://doi.org/10.1016/j.trc.2021.103357

[21] S. Kumar, P. Kaur, A. Gosain, A comprehensive survey on ensemble methods, in: 2022 IEEE 7th International conference for Convergence in Technology (I2CT), 2022, pp. 1–7. https://doi.org/10.1109/I2CT54291.2022.9825269

[22] N.R. Swanson, H. White, A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks, Rev. Econ. Statist. 79 (4) (1997) 540–550.

[23] B. Billah, M.L. King, R.D. Snyder, A.B. Koehler, Exponential smoothing model selection for forecasting, Int. J. Forecast. 22 (2) (2006) 239–247. https://doi.org/10.1016/j.ijforecast.2005.08.002

[24] J. Liu, Z. Yu, H. Zuo, R. Fu, X. Feng, Multi-stage residual life prediction of aero-engine based on real-time clustering and combined prediction model, Reliab. Eng. Syst. Safe. 225 (2022) 108624. https://doi.org/10.1016/j.ress.2022.108624

[25] P. Bertail, P. Doukhan, P. Soulier, Dependence in Probability and Statistics, Springer, New York, 2006.

[26] A. Wald, Sequential tests of statistical hypotheses, Annal. Math. Statist. 16 (2) (1945) 117–186. http://www.jstor.org/stable/2235829.

[27] J.-L. Vivancos, R.A. Buswell, P. Cosar-Jorda, C. Aparicio-Fernández, The application of quality control charts for identifying changes in time-series home energy data, Energy Build. 215 (2020) 109841. https://doi.org/10.1016/j.enbuild.2020.109841

[28] A. Ramdas, P. Grünwald, V. Vovk, G. Shafer, Game-theoretic statistics and safe anytime-valid inference, Statist. Sci. 38 (4) (2023) 576–601.

[29] G. Shafer, A. Shen, N. Vereshchagin, V. Vovk, Test martingales, bayes factors and p-values, Statist. Sci. 26 (1) (2011) 84–101. https://doi.org/10.1214/10-STS347

[30] A. Henzi, J.F. Ziegel, Valid sequential inference on probability forecast performance, Biometrika 109 (3) (2022) 647–663.

[31] Y.J. Choe, A. Ramdas, Comparing sequential forecasters, Oper. Res. 72 (4) (2024) 1368–1387. 2110.00115

[32] J. Coignard, M. Janvier, V. Debusschere, G. Moreau, S. Chollet, R. Caire, Evaluating forecasting methods in the context of local energy communities, Int. J. Electr. Power Energy Syst. 131 (2021) 106956. https://doi.org/10.1016/j.ijepes.2021.106956

[33] A. Klenke, Probability Theory : a Comprehensive Course, Springer International Publishing, 3rd edition, 2020.

[34] G. Shafer, Testing by betting: a strategy for statistical and scientific communication, J. Roy. Statist. Soc. Ser. A: Statist. Soc. 184 (2) (2021) 407–431. https://academic.oup.com/jrsssa/article-pdf/184/2/407/49325712/jrsssa_184_2_407.pdf

[35] V. Vovk, R. Wang, E-values: calibration, combination and applications, Annal. Statist. 49 (3) (2021) 1736–1754.

[36] R.G. Miller, Simultaneous Statistical Inference, Springer, 2nd edition, 1981.

[37] P. Grünwald, R. de Heide, W. Koolen, Safe testing. Discussion paper, J. Roy. Statist. Soc. Ser. B: Statist. Methodol. (Accepted/In press - 24 Jan 2024). Accepted author version available under https://research.utwente.nl/files/353763913/Preprint-Grunwald-24-Jan-2024.pdf.

[38] J.P. Romano, A.M. Shaikh, M. Wolf, et al., Multiple Testing, The New Palgrave Dictionary of Economics, 9185–9189 (2018), Palgrave Macmillan UK, London, 978-1-349-95189-5, https://doi.org/10.1057/978-1-349-95189-5_2914

[39] B. De Finetti, Sulle stratificazioni convesse, Annali di Matematica Pura ed Applicata 30 (1949) 173–183.

[40] T. Gneiting, Making and evaluating point forecasts, J. Am. Stat. Assoc. 494 (106) (2011) 746–762 https://doi.org/10.1198/jasa.2011.r10138

[41] T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc. 102 (477) (2007) 359–378. http://www.jstor.org/stable/27639845.

[42] S.R. Howard, A. Ramdas, J. McAuliffe, J. Sekhon, Time-uniform, nonparametric, nonasymptotic confidence sequences, Annal. Statist. 49 (2) (2021) 1055–1080 https://doi.org/10.1214/20-AOS1991

[43] S.R. Howard, A. Ramdas, J. McAuliffe, J. Sekhon, Time-uniform Chernoff bounds via nonnegative supermartingales, Probab. Surv. (17) (2020) 257–317. arXiv:1808.03204. https://doi.org/10.1214/18-PS321

[44] C. Gilbert, J. Browell, B. Stephen, Probabilistic load forecasting for the low voltage network: forecast fusion and daily peaks, Sustain. Energy Grid. Netw. 34 (2023) 100998. https://doi.org/10.1016/j.segan.2023.100998

[45] D. Musikhina, J. Seidemann, S. Feilmeier, 2021. EMSIG: Energy Management System, Data, available at:, https://openenergyplatform.org/dataedit/view/demand/emsig_energy_data_by_ems (2024).

[46] S.N. Wood, Generalized Additive Models: an Introduction with R, Chapman and Hall/CRC, 2017.

[47] S.N. Wood, Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, J. Roy. Statist. Soc. Ser. B: Statist. Methodol. 73 (1) (2011) 3–36.

[48] D.J. Gauthier, E. Bollt, A. Griffith, W.A.S. Barbosa, Next generation reservoir computing, Nat. Commun. 12 (1) (2021) 5564. https://doi.org/10.1038/s41467-021-25801-2

[49] I. Ratas, K. Pyragas, Application of next-generation reservoir computing for predicting chaotic systems from partial observations, Phys. Rev. E 109 (2024) 064215. https://doi.org/10.1103/PhysRevE.109.064215

[50] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neur. Comput. 9 (8) (1997) 1735–1780.

[51] T. Hong, S. Fan, Probabilistic electric load forecasting: a tutorial review, Int. J. Forecast. 32 (3) (2016) 914–938. https://doi.org/10.1016/j.ijforecast.2015.11.011

[52] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-Generation Hyperparameter Optimization Framework, in: The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.

[53] S. Ryu, H. Choi, H. Lee, H. Kim, V.W.S. Wong, Residential Load Profile Clustering via Deep Convolutional Autoencoder, in: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), 2018, pp. 1–6. https://doi.org/10.1109/SmartGridComm.2018.8587454

[54] L. Held, M. Ott, On p-values and bayes factors, Annu. Rev. Statist. Appl. 5 (Volume 5, 2018) (2018) 393–419. https://doi.org/10.1146/annurev-statistics-031017-100307

[55] E. Kaufmann, W.M. Koolen, Mixture martingales revisited with applications to sequential tests and confidence intervals, J. Mach. Learn. Res. 22 (246) (2021) 1–44.

[56] A.P. Dawid, Present position and potential developments: some personal views statistical theory the prequential approach, J. Roy. Statist. Soc.: Ser. A (General) 147 (2) (1984) 278–290.

[57] Q. Wang, R. Wang, J. Ziegel, E-backtesting, 2024. arXiv:2209.00991.

[58] Q. Wang, R. Wang, J. Ziegel, Simulation and Data Analysis for E-backtesting, 2023. https://doi.org/10.2139/ssrn.4346325

[59] T. Gneiting, M. Katzfuss, Probabilistic forecasting, Annu. Rev. Statist. Appl. 1 (1) (2014) 125–151 https://doi.org/10.1146/annurev-statistics-062713-085831