

Secondary Publication



Hopf, Konstantin; Hartstang, Hannah; Staake, Thorsten Robert

Meta-Regression Analysis of Errors in Short-Term Electricity Load Forecasting

Date of secondary publication: 24.10.2023

Accepted Manuscript (Postprint), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-913046

Primary publication

Hopf, Konstantin; Hartstang, Hannah; Staake, Thorsten Robert: Meta-Regression Analysis of Errors in Short-Term Electricity Load Forecasting. In: Companion Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy) ; e-Energy Workshop 2023: International Workshop on Energy Data and Analytics ; Orlando, FL. ACM digital library S. 32-39, 2023. DOI: 10.1145/3599733.3600248

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holder(s).

This document is made available with all rights reserved.

Meta-Regression Analysis of Errors in Short-Term Electricity Load Forecasting

Konstantin Hopf
University of Bamberg
Bamberg, Germany
konstantin.hopf@uni-bamberg.de

Hannah Hartstang
University of Bamberg
Bamberg, Germany
hannah.hartstang@gmx.de

Thorsten Staake
University of Bamberg
Bamberg, Germany
thorsten.staake@uni-bamberg.de

ABSTRACT

Forecasting electricity demand plays a critical role in ensuring reliable and cost-efficient operation of the electricity supply. With the global transition to distributed renewable energy sources and the electrification of heating and transportation, accurate load forecasts become even more important. While numerous empirical studies and a handful of review articles exist, there is surprisingly little quantitative analysis of the literature, most notably none that identifies the impact of factors on forecasting performance across the entirety of empirical studies. In this article, we therefore present a Meta-Regression Analysis (MRA) that examines factors that influence the accuracy of short-term electricity load forecasts. We use data from 421 forecast models published in 59 studies. While the grid level (esp. individual vs. aggregated vs. system), the forecast granularity, and the algorithms used seem to have a significant impact on the MAPE, bibliometric data, dataset sizes, and prediction horizon show no significant effect. We found the LSTM approach and a combination of neural networks with other approaches to be the best forecasting methods. The results help practitioners and researchers to make meaningful model choices. Yet, this paper calls for further MRA in the field of load forecasting to close the blind spots in research and practice of load forecasting.

KEYWORDS

Electricity Demand Forecast, Short-Term Forecasting, Meta Regression Analysis (MRA), Mean Absolute Percentage Error (MAPE)

©Hopf, Hartstang, Staake (2023). This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive version was published in the *Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy '23)*, June 20–23, 2023, Orlando, FL, USA, <https://doi.org/10.1145/3599733.3600248>.

1 INTRODUCTION

Accurate forecasting of electricity demand is an important success factor for utilities, and there is reason to believe that such forecasts will become even more important in the future: The electrification of residential heating and the adoption of electric vehicles will increase both, volatility of demand and utilization of the distribution grid [5]. As a result, the safety margins of existing assets will decrease while energy costs at peak load times will rise. Load control—both centralized, e.g., by grid operators, and decentralized by local agents—will benefit from accurate demand forecasts, as will utilities' attempts to schedule production and hedge demand through forward contracts.

As a consequence, there is an extensive literature on methods and models for electric load forecasting [15], which has been summarized in recent review papers [e.g., 12, 22, 26]. The academic discourse is fueled by the continued global deployment of advanced metering infrastructures that makes an increasing amount of consumption data available at higher temporal resolution. Thus, such data have attracted significant research interest to investigate the use of smart meter data for load forecasting and to enable electricity forecasting at different grid levels and for different time scales [31].

The existing literature reviews on short-term electricity load forecasting provide good overviews of the large number of empirical studies. They show which computational methods are used to forecast future electricity demand for different forecast horizons and grid levels. The reviews also identify blind spots in existing research and outline research agendas. Yet, the existing review studies are primarily qualitative.

Quantitative meta-reviews, by contrast, aim to build a mean effect size from comparable, independent individual studies. Thus, quantitative meta-reviews allow for more reliable results than single empirical studies [21]. They also enable the identification of parameters that explain variances and heterogeneity of effect sizes in study results when sufficient data are available [6, 24]. Such quantitative meta-reviews are particularly helpful for practitioners who strive to operationalize forecasts for specific situations and want to draw on evidence from a complete research field [6]. They also help research to judge the robustness of existing approaches, recognize patterns within working solutions, and identify outliers that might be especially promising or questionable.

Our study seeks to extend the previous reviews with an inductive statistical analysis, as we carry out a Meta Regression Analysis (MRA) for short-term electricity load forecasts. By short-term time horizon, we mean load forecasts with up to one week ahead [3, 12, 22]. We thereby aim to explain factors—across a large sample of individual studies examining electric load forecasting—that lead to high or low quality short-term electric load forecasts.

Our article starts with an overview of recent electric load forecasting review studies, describes the method, and our analytical results. We conclude with an interpretation and outline future directions for research and practice.

2 BACKGROUND

The field of electricity load forecasting is comprehensive and stays in connection to other fields of energy forecasting [15]. We found several review articles that were published in the last five years (see Table 1) and provide a summarized overview of the field. Similar to trends that Hong et al. [15] identify for the broader field of energy forecasting, the literature on electricity load forecasting heavily

uses developments in the field of Machine Learning (ML). Another observation that holds for the fields of energy and electricity load forecasting alike is that studies primarily focus on forecasts at the system level or the transmission grid and (through the proliferation of smart meter data in recent years) also on the household level [12, 15]. Other grid levels have not been in the focus yet.

Table 1: Literature reviews on electricity forecasting in the last five years

Ref.	Year	Journal	Dep. variable	Focus
[12]*	2021	Appl. En.	load	low voltage grid
[1]	2021	Ren. and Sust. En. Rev.	load + prod.	no restriction
[30]	2021	Energies	load	manufacturing
[22]*	2020	J. El. Sys. and IT	load	no restriction
[26]*	2020	En. and Buildings	load	buildings
[29]	2020	Entropy	load + prod.	no restriction
[8]	2020	Intl. J. of En. Res.	load + prod.	no restriction
[3]	2019	Sust. Cities and Society	load + prod.	buildings
[31]*	2018	IEEE Tran. on Smart Grid	load	no restriction

The review articles describe the landscape of electricity load forecasting, list algorithmic approaches, datasets, various forecasting horizons, and grid levels. They also problematize implicit field assumptions, point out limitations in the field, and identify future research directions. Nevertheless, the review articles we found are primarily qualitative summaries. If the reviews include quantitative analyses, the evaluations of forecasting models focus on bibliometrics (e.g., publication date, journal) or analyze the prediction models in a descriptive way (e.g., frequency of algorithm categories). Only Vivas et al. [29] investigate the relationship of broad algorithm classes (statistical vs. machine learning vs. hybrid) and data granularity on model performance, but are not examining influence factors on prediction performance with inductive statistics.

Thus, for practitioners that want to operationalize forecasts, it is difficult to decide which algorithmic approach is suitable for a given application when such aggregated knowledge does not exist.

3 METHOD

The approach of MRA goes back to Glass [10] who proposed the method in 1976 as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" [10, p. 3]. MRAs find wide application in many fields such as medicine, psychology, and economics [21, 24]. In research related to the energy domain MRAs exist, for example, on water demand [23] and energy prices [11].

MRAs typically follow the following three steps [25]: First, define the research question and the effect size (as the core criterion of interest), second, literature search and coding, and third, meta-regression modeling. Our description below follows these three steps.

3.1 Review focus and effect size

Our focus lies on the predictive quality of short-term electricity forecasts (up to week-ahead) using point-estimates. To evaluate such forecasts, several performance metrics exist [17]. Absolute error metrics, like Mean Absolute Error (MAE) or Root Mean Square

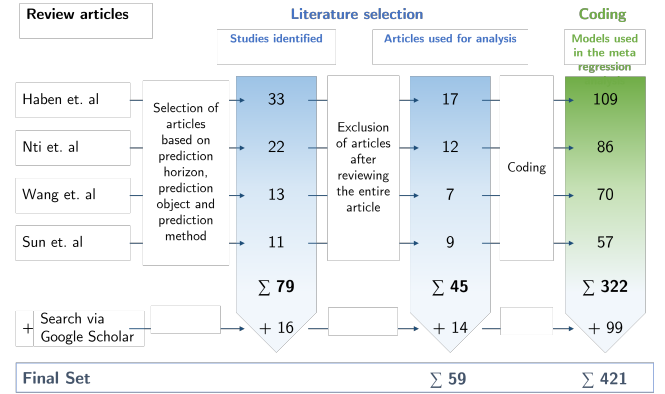


Figure 1: Literature selection and coding process

Error (RMSE) are not helpful as they are scaled in the dimension of the input data, thus, do not allow comparison across studies.

An error measure that is scale-independent and allows comparisons across studies is the Mean Absolute Percentage Error (MAPE) metric. As an alternative, the Normalized Root Mean Square Error (NRMSE) could also be used, but we found that MAPE is more frequently reported (77.2% of studies in our sample reported MAPE and only 12.7% NRMSE). Thus, we selected MAPE as the effect size.

3.2 Literature selection

As MRAs aim to provide a comprehensive overview to a field, the selection of the sample of empirical studies included into the analysis is crucial and should be consciously made [24]. We decided to use empirical studies that were mentioned in four review articles on electricity load forecasting that appeared in the last five years. All review articles report a systematic review process with clear selection criteria [4, 14]. Each review article covers a slightly different review focus, which increases the breath of our sample.

In detail, we use the review study by Nti et al. [22], which focuses on electricity load forecasting in general, the review by Haben et al. [12], which focuses on short-term electricity forecasting in the low-voltage grid, the review of Sun et al. [26], which focuses on forecasts electricity use of buildings, and the review by Wang et al. [31], which focuses on the use of smart meter data. We mark the used studies with an asterisk in Table 1.

Our selection of empirical studies from these review papers followed three steps, as we illustrate in Figure 1. We first screened each review paper for empirical papers that investigated a short-term prediction horizon (up to week-ahead), used electric load as a dependent variable and obtained point estimates (we excluded studies on probabilistic forecasting). We made this initial selection of articles based on the information presented in the review article. This initial screening led us to 79 references.

Second, we had to exclude studies after our in-depth reading with the following reasons: As we use MAPE as an effect size metric, we must exclude articles that do not report MAPE values or do not clearly document the actual numbers (e.g., that just show bars in graphs without actual numbers). In addition, we excluded review

studies or secondary studies that just compared results of other primary studies. Furthermore, we excluded studies that did not allow any conclusions about the sample size of the data used for training and evaluation. In total, 45 articles remained after completing this second article screening step.

As a third step, we conducted a limited structured search of articles via Google Scholar, in order to complete the sample for our MRA. We used the search terms "short-term", "forecast*", "MAPE", "energy", "load", and "demand" and reviewed the hits on the first pages according to the first and second step described above. This search yielded to additional 16 articles, of which we included 14 in our analysis, leading us to a sample of 59 articles.

3.3 Coding

During our in-depth reading of the articles, we extracted all models that the studies report together with all relevant information for our MRA. For this, we use a systematic coding procedure [18], following the coding guide that we describe in the remaining section. We list the resulting variables in Table 2 and in Table 3.

Effect size and sample sizes: We coded the MAPE values in percent, the number of observation points as integer, and transformed the timespans of the data reported in the papers as the number of days.

Forecast horizon: We coded the forecast horizon h in a categorical and in a numeric variable. For the categorical, we differentiate between four ordered classes $h \leq 1\text{hourahead}$, $1\text{hourahead} < h < 1\text{dayahead}$, $h = \text{dayahead}$, $h > \text{dayahead}$. For the numeric coding, we gathered the number of half-hour steps.

Forecast granularity: We also coded the forecast granularity, which is the resolution in which the forecast is computed. To harmonize this across the different studies, we express the granularity relative to a half-hour step, i.e., hourly granularity would be 2 and quarterly would be 0.5.

Model category: Finally, we classified the models in the studies into one of 17 categories of algorithmic approaches. For this, we created a classification scheme based on schemata used in the considered review papers [12, 22, 26, 29] but also others [7, 13]. A common division of forecasting models makes a distinction between statistical, ML, and hybrid models. We further differentiate these three groups into different algorithmic approaches.

Among *statistical methods*, we differentiate between *time series* (including AR(X), NARX, MA, ARMA(X), ARIMA(X), SARIMA(X), state space, and spatio-temporal models), *regression* (including linear and multilinear regressions) and *exponential smoothing* methods (e.g., Holt-Winters models).

Table 2: Numeric variables

Variable	Description	Mean	Std. Dev.
MAPE	Percent	16.07	23.92
Horizon_Num	In half-hour steps	49.7	60.1
Granularity_Num	In half-hourly values	7.83	15.76
Year	Year of publication	2017.39	2.64
N_days	Days for model training and test	460.2	486.7
N_obs	No. of observation points	334	3,948

In the category of *ML models*, Neuronal Networks (NNs) are the most frequent algorithm category. We split this category into *Shallow_NN* and *Deep_NN*, where *Shallow_NN* are those having only one hidden layer and *Deep_NN* are those having more than one hidden layer. To keep the number of categories manageable, we just differentiate between network architectures, in particular, architectures for time-series data, i.e., Recurrent Neural Networks (RNNs) and Long Short Term Memorys (LSTMs) (which is an advancement of the RNN approach). We combined Support Vector Machine (SVM) and Support Vector Regression (SVR) into one category. Other machine learning models include *Boltzmann* machine models, *k nearest Neighbor (kNN)*, *fuzzy logic*, and *ensemble and other* models (e.g., autoencoders, genetic algorithms, gradient boosting, random forest, and decision tree algorithms).

Several studies propose *hybrid models* by combining an algorithm from one of the considered classes with another (e.g., a linear regression model). We consider two classes of hybrid models, those with NNs (*Hybrid_NN*) and those with others (*Hybrid_various*).

Finally, multiple studies use *benchmark estimators*, which do not transform the data in a sophisticated way. These simple models are grouped in their own class.

To get a better understanding of the use of the different model categories in the studies, Figure 2 shows in the number of studies that use the different algorithm classes over time. Looking at the NN-based approaches, we notice that before 2018 mainly *Shallow_NN* were used. After 2018, the models with multiple hidden layers dominated. This indicates an evolution in research on NNs. LSTMs are gaining momentum over time and are the most common

Table 3: Categorical variables

Variable	Categories	No. Models
Level	Individual (household, building, ...)	146 (34.7%)
	Aggregated (sum of multiple entities)	105 (24.9%)
	Substation (transformer, grid zone, ...)	23 (5.4%)
	System	147 (34.9%)
Horizon_Cat	$h \leq 1\text{hourahead}$	64 (15.2%)
	$1\text{hourahead} < h < 1\text{dayahead}$	29 (6.9%)
	$h = \text{dayahead}$	308 (73.2%)
	$h > \text{dayahead}$	20 (4.7%)
Model_Cat	Stat: Time Series	52 (12.3%)
	Stat: Regression	42 (10.0%)
	Stat: Exponential Smoothing	16 (3.8%)
	ML: Shallow_NN	82 (19.5%)
	ML: Deep_NN	50 (11.9%)
	ML: RNN	13 (3.1%)
	ML: LSTM	29 (6.9%)
	ML: Boltzmann	3 (0.7%)
	ML: SVM_SVR	50 (11.9%)
	ML: kNN	7 (1.7%)
	ML: Fuzzy_Logic	14 (3.3%)
	ML: Ensemble and other	15 (3.6%)
	Hybrid_NN	22 (5.2%)
	Hybrid_various	2 (0.47%)
Benchmark	24 (5.7%)	
Study_ID	<i>Unique per study</i>	421 (100%)

approach used in the studies published in 2021. Time series and benchmark models are used fairly regularly over the years and most often serve as reference models in the studies to compare. We also find that some approaches are used only very rarely, for example, fuzzy logic, ensemble methods, kNNs, Boltzmann Machine, Genetic Algorithms, and Autoencoders. We therefore group these infrequent model categories together for the following analyses.

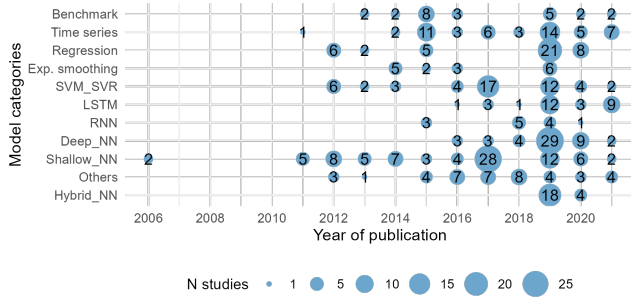


Figure 2: Illustration of the use of algorithmic categories in the studies over time

3.4 Statistical analysis

For the statistical analysis of our MRA, we rely on Ordinary Least Squares (OLS) and Weighted Least Squares (WLS) regression. Guidelines for conducting MRAs point to the problems of heteroscedasticity, reliability and interdependence between examined factors [21, 23], which we address through the following approaches.

First, to mitigate the problem of heteroscedasticity, we use robust standard errors [32, 33]. Second, to include an estimate of the study reliability, we weight the effect sizes of the single empirical studies. Meta-analyses frequently use the variance of the effect sizes as a weighting factor, given that effect sizes with smaller variances are considered as more reliable and should be weighted more heavily in a MRA. In the field of electricity forecasting, the variance of error metrics are, however, usually not reported. Therefore, we estimate the reliability of a study using the sample size [21]. As we have time-series data available, and ideally time series data from multiple observation points, we compute

$$SampleSize = N_{days} * N_{obs} \quad (1)$$

As some studies have very large samples, we use the logarithm to lower the influence of very large sample sizes and scale the weighting factors using max-normalization.

Third, we evaluate interdependence between examined influence factors and the effect size. One reason can be that multiple primary studies use the same data set, which is partly the case for electricity forecasting [12]. Another reason is that multiple effect sizes may be reported from a single study. Observable common effects, such as the common data set, can be accounted for using regressors. To account for study-specific influences, we use fixed effects models in that we estimate a regression intercept per study [2, 21].

4 RESULTS

4.1 Study-specific parameters

For the models that examine study-specific characteristics, we computed simple linear models using OLS to test the correlation between MAPE and one variable as regressor each. First, we could not find an influence of the year of publication on the error values ($R^2 = 0.01$, $F(421) = 2.36$, $p = .1612$). Second, we tested the influence of the size of the data set, that is, the observation period in days and the number of observation points (e.g., meters or households). While the regression models found statistically significant effects of the size of the data, the effect sizes are very small and the variances explained by the study parameters are low for the observation period ($R^2 = .000$, $F(421) = 0.39$, $p < .001$) and the number of observation points ($R^2 = .008$, $F(421) = 3.48$, $p < .05$).

4.2 Grid levels

As a second analysis, we examined the influence of the grid level on the forecasting error. We encoded the grid level as dummy variables, $d_{ind} = 1$ if the forecast was made on an individual level (e.g., households, buildings), $d_{sub} = 1$ if the forecast was obtained for the substation-level, and $d_{aggr} = 1$ if the forecast was made by aggregating time-series from lower grid levels. The case that the forecast targeted the system level is represented as response state (i.e., all dummy variables are zero) because this is the most frequent grid level across all studies. We use a fixed-effects model considering an intercept for each study (represented by the study ID_i , $i \in 1, \dots, 59$) with the coefficient β_{0i} . We estimated the model using a WLS estimation using $\log(SampleSize)$ as a weighting factor (see Equation 1) and used robust standard errors to address heterogeneity [32, 33] with the following model specification:

$$y_i = \beta_{0i} * ID_i + \beta_1 * d_{ind} + \beta_2 * d_{sub} + \beta_3 * d_{aggr} + \epsilon_i \quad (2)$$

Table 4 shows the WLS model estimates. Due to the study-specific intercept, the proportion of explained variance is quite high (this holds also true for the following models). Accordingly, we focus our interpretation on the main effects and their significant difference from zero.

As the descriptive plot in Figure 3 shows, the studies with individual-level prediction have a large variance and report significantly higher error values than the models with other grid levels. Compared to prediction for a single time-series of a complete power

Table 4: MAPE estimates for different grid levels, using the individual household as a baseline (model 1)

Model 1	
Individual	25.16 (4.16)***
Aggregated	-0.62 (3.15)
Substation	0.18 (2.12)
R^2	0.72
Adj. R^2	0.67
Num. obs.	421
F statistic	14.79

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

system (reference category), our model estimates that the forecast errors on individual level are 25.16 percentage points higher. If predictions are made with aggregated data, this seems to lead to better predictions (in 0.62 percentage points lower error), yet the difference is not statistically significant. The lacking significant level for forecasts at the aggregated and the secondary grid level (i.e., substation) might be due to the small number of studies that consider this level of forecast.

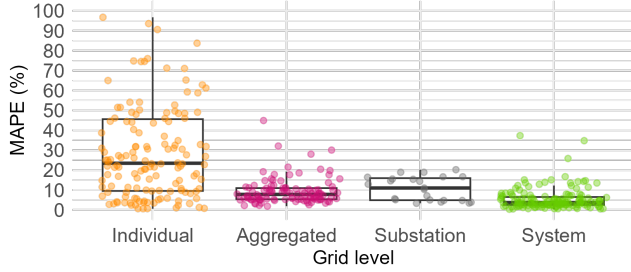


Figure 3: Combined scatter- and boxplot for MAPE values across different grid levels

Reasons for the high relative errors in individual-level studies may be that individual load curves might be harder to predict than aggregated ones. Another explanation is that individual load curves often have times with small load, which increases the MAPE. To investigate this issue further, other error metrics that give less weight to small consumption values may need to be included in the analysis.

4.3 Time horizon

Similar to the previous analysis, we tested if the forecasting horizon has an impact on the errors. Only few models had a horizon between hour- and day-ahead (29 models), and even fewer have a forecasting horizon of more than day-ahead (20 models). Thus, we excluded these models for this analysis. For the remaining data (372 models), we encoded the day-ahead as $d_{dayahead} = 1$ and the hour-ahead forecasts as $d_{dayahead} = 0$ and estimate the following regression model:

$$y_i = \beta_{0i} * ID_i + \beta_1 * d_{dayahead} \epsilon_i \quad (3)$$

Given that the dummy encoding omits ranking information and we left out several models, we used an alternative model formulation with a metric variable. We defined a variable that expresses the forecasting horizon in the number of 30-minute time steps (i.e., an hourly forecast has $horizon_timesteps = 2$ and a 24h forecast $horizon_timesteps = 48$). This approach follows earlier meta-reviews in the field of electricity [23] and traffic forecasting [27].

$$y_i = \beta_{0i} * ID_i + \beta_1 * horizon_timesteps + \epsilon_i \quad (4)$$

Table 5 shows the WLS model estimates. From both models, it appears that there is no significant influence of the forecast horizon on model quality. This may be because the effect is confounded by algorithm choice or practical relevance over time. The relatively high R^2 values result from the study-specific intercept in our fixed-effects model.

Table 5: MAPE estimates for different time horizons, using dummy (model 2) and using numeric encoding (model 3)

	Model 2	Model 3
Dayahead	0.63 (5.36)	
Horizon_Num		0.01 (0.05)
R^2	0.67	0.68
Adj. R^2	0.61	0.63
Num. obs.	372	421
F statistic	10.91	12.73

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

4.4 Forecast granularity

As a fourth analysis, we examine the influence of the forecast granularity on the size of the error, using the model:

$$y_i = \beta_{0i} * ID_i + \beta_1 * granularity + \epsilon_i \quad (5)$$

The regression results in Table 6 show that granularity has a negative and significant effect on the magnitude of the prediction error ($R^2 = .68$, $F(60, 361) = 12.9$, $p < .001$). This means that larger time steps of the forecasts lead to lower relative prediction errors. For every half hour that the forecast granularity increases, the error decreases by 0.39 percentage points.

Table 6: MAPE estimates for different data granularities

	Model 4
Granularity_Num	-0.39 (0.14)**
R^2	0.68
Adj. R^2	0.63
Num. obs.	421
F statistic	12.90

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

4.5 Algorithm category

Finally, we investigate the influence of the model category on the forecasting error. As algorithmic innovations aim to improve forecasts, we expect a strong influence of the model category on the forecast errors [19, 29].

We consider the algorithm class as a dummy-encoded variable $Model_Category_j$ for each model category, being 1 if the class is used, 0 otherwise. We use *regression models* as the response category (because they have the worst forecasting performance in the studied models), meaning that all $Model_Category_j = 0$. Given that several model categories are very infrequent, we used ten model categories as shown in Table 7, thus $1 \leq j \leq 10$.

$$y_i = \beta_{0i} * ID_i + \beta_j * Model_Category_{ij} + \epsilon_i \quad (6)$$

The results of the regression analysis in Table 7 (*Overall*) show that all model categories except the *Shallow_NNs* seem to produce better results than the *regression* (which is the response category), even the benchmark models. Yet, only the coefficients of the model

Table 7: MAPE results for algorithm category (overall and across different grid levels)

	Overall	Individual	Aggregated	System
Shallow_NN	1.51 (4.28)	3.27 (11.41)	-0.76 (2.53)	-3.57 (1.27)**
Deep_NN	-8.74 (4.48)	-15.38 (13.29)	-8.26 (2.55)**	4.80 (1.24)***
Time_Series	-0.56 (4.25)	-2.69 (12.66)	-2.96 (2.54)	-2.78 (1.20)*
RNN	-7.16 (6.24)	-16.83 (15.89)	-7.46 (3.20)*	-1.16 (2.17)
LSTM	-15.85 (5.26)**	-27.54 (11.61)*	-8.66 (2.96)**	
SVM_SVR	-3.69 (4.98)	-1.78 (13.26)	-3.60 (2.97)	-4.86 (1.30)***
Benchmark	-3.81 (5.37)	-19.94 (14.21)	1.52 (2.82)	1.40 (2.22)
Exponential_Smoothing	-1.34 (5.54)	-6.61 (15.86)	-3.68 (2.83)	5.18 (2.05)*
Hybrid_NN	-15.65 (5.89)**	-34.89 (15.77)*	-8.37 (2.94)**	-1.26 (5.00)
Others	-0.57 (4.81)	-0.37 (11.74)	-7.53 (2.82)**	-4.27 (1.49)**
R^2	0.69	0.77	0.87	0.88
Adj. R^2	0.63	0.69	0.82	0.85
Num. obs.	421	146	105	147
F statistic	11.56	9.81	18.64	29.10

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; \cdot $p < 0.1$

categories *Deep_NN*, *LSTM*, and *Hybrid_NN* have a significant effect on the magnitude of the forecast error overall.

In our first analysis, we have found that the grid level has a significant influence on the model results. Therefore, we also analyzed the model categories with data subsets of studies focusing on an individual, aggregated, and system level (the substation category was too infrequent that we could compute the model). For the system level, there is only one study in our sample using the *LSTMs* approach, thus, our model cannot estimate a coefficient due to perfect collinearity in this case.

We see that *LSTMs* and *Hybrid_NNs* show the lowest prediction errors and also have significant effects in the *Individual* and *Aggregated* subsample. Thus, we conclude that these two approaches lead to the best forecasting results in the sample of analyzed studies for the two grid levels.

For the grid level, the performance figures are quite different, suggesting that *SVM_SVR*, *Shallow_NN* and *Other* approaches lead to lower results. Yet, the otherwise strong category of *LSTM* has been left out of the calculation because of just a single study. Future research should, thus, investigate the performance of algorithm categories on a system level including further and, in particular, more recent studies.

5 DISCUSSION

The in-depth analysis of error metrics using a MRA helps to identify parameters that explain variances and heterogeneity of effect sizes in a large number of empirical studies [6, 21, 24]. For the field of short-term electricity load forecasting, MRAs can help to identify parameters, algorithms, and situations that foster smaller forecast errors.

Our analysis showed, for example, that the grid level (esp. individual vs. aggregated vs. system), the forecast granularity, and the algorithms used have a significant impact on the MAPE. We also found that the *LSTM* approach and a combination of NNs with other methods were the best forecasting methods for the individual and aggregated forecasts. For system level forecasts, *SVM*, *SVR*,

Shallow NNs, and other approaches seem to perform best in our sample. In contrast, the year of publication, dataset size, and prediction horizon had no significant effect on prediction performance in our sample of studies.

The results help practitioners to operationalize forecasting models for specific applications, drawing on the aggregated findings of 59 empirical studies. For researchers, the results help to assess the robustness of approaches they suggest, identify patterns and blind spots in the variety of existing solutions, and identify outliers that may be particularly promising or questionable.

5.1 Future work

The analysis we presented in this paper is promising and should be a call to the load forecasting research field to look more closely at MRAs. Many aspects could not be addressed in this study and thus require follow-up investigations.

First, the sample of studies would need to be expanded to include older studies to better reflect trends over time in this area. Beyond that, a broader sample would strengthen the analysis. Extending the sample would also allow more detailed insights in subgroup analyses, for example, if some algorithms are better-suited for certain grid levels or data sources than others.

Second, the field of load forecasting is constantly evolving and new approaches, such as the Transformers architecture [9, 20, 28], could not be included in our evaluation so far.

Third, we also did not control for the datasets used. Even though many studies use datasets that are not public, there are some datasets that are used very often [12]. This may bias the analysis. An in-depth analysis in terms of datasets (also the statistical properties of the datasets used in the empirical studies) could provide further insight into, for example, how larger training datasets and high-resolution data have an impact on predictive performance.

Fourth, an investigation of the influence of different data sources and features used (e.g., weather, geospatial information), also with a focus on open data [16], on forecast performance would also be an exciting extension.

Finally, future research should also apply MRAs to probabilistic load forecasts. We could not include such studies because probabilistic load forecasts are evaluated with other error metrics and thus a different dependent variable would be necessary.

5.2 Limitations

A limitation of our analysis is the use of the MAPE metric, which is highly dependent on actual consumption in the evaluation. We would have liked to use a more reliable quality metric, such as NRMSE but other metrics that allow comparison between studies are rarely reported. Future research in the prediction literature should therefore report more quality metrics that allow quantitative comparative analysis.

6 CONCLUSION

The research literature on short-term electricity load forecasting is extensive, and previous survey articles summarize the field descriptively. To our knowledge, our analysis is the first MRA to quantitatively examine factors influencing forecast quality.

To close this gap, we have analyzed the prediction errors of 421 models published in 59 studies that were mentioned in recent review articles in the field of short-term electricity load forecasting. Our statistical MRA could find statistically significant influences of (i) the grid level (individual, aggregated, and system), (ii) the forecast granularity, and (iii) the algorithms used (particularly good approaches are LSTM and Hybrid_NN on the individual and the aggregated level, while SVM, shallow NN, and other ML approaches perform best for grid-level forecasts) on the MAPE reported in these studies. We did not find an influence of the study characteristics (year of publication, dataset size) or the time horizon of the forecast on the MAPE.

Although short-term load forecasting offers powerful tools with acceptable forecast error metrics, the development of new forecasting methods remains a major challenge. In the future, the influences of the energy transition will have a greater impact on all energy consumption sectors. For example, heat pumps and electric vehicles will proliferate, causing additional loads. Weather influences on electricity consumption will change, caused by environmental change and increase power-to-heat appliances. Moreover, as some energy providers experiment with variable tariffs and customers invest in home energy management systems that can optimize for market prices in addition to self-consumption, demand profiles will incorporate market feedback that will alter electricity demand profiles further. Future research, thus, has to include novel predictors, like short term price elasticity or technical aspects like the common ripple control in some countries, which, because of the control, could improve the prediction quality.

ACKNOWLEDGMENTS

We thank the Bavarian Ministry of Economic Affairs, Regional Development and Energy for their financial support of the project "DigiSWM" (DIK-2103-0014), as part of which the study was carried out.

REFERENCES

- [1] Sheraz Aslam, Herodotos Herodotou, Syed Muhammad Mohsin, Nadeem Javaid, Nouman Ashraf, and Shahzad Aslam. 2021. A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. *Renewable and Sustainable Energy Reviews* 144 (2021), 1–55.
- [2] Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods* 1, 2 (2010), 97–111.
- [3] Mathieu Bourdeau, Xiao qiang Zhai, Elyes Nefzaoui, Xiaofeng Guo, and Patrice Chatellier. 2019. Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustainable Cities and Society* 48 (2019), 1–27.
- [4] Iain Chalmers, Larry V. Hedges, and Harris Cooper. 2002. A brief history of research synthesis. *Evaluation & the health professions* 25, 1 (2002), 12–37.
- [5] Spyridon Chapaloglou, Athanasios Nesiadis, Petros Iliadis, Konstantinos Atsonios, Nikos Nikolopoulos, Panagiotis Grammelis, Christos Yiakopoulos, Ioannis Antoniadis, and Emmanuel Kakaras. 2019. Smart energy management algorithm for load smoothing and peak shaving based on load forecasting of an island's power system. *Applied Energy* 238 (2019), 627–642.
- [6] Harris Cooper, Larry V. Hedges, and Jeffrey C. Valentine. 2009. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation. <https://www.jstor.org/stable/10.7758/9781610441384>
- [7] Kumar Biswajit Debnath and Monjur Mourshed. 2018. Forecasting methods in energy planning models. *Renewable and Sustainable Energy Reviews* 88 (2018), 297–325.
- [8] Jayanthi Devaraj, Rajvikram Madurai Elavarasan, G. M. Shafiqullah, Taskin Jamal, and Irfan Khan. 2021. A holistic review on energy forecasting using big data and deep learning models. *International journal of energy research* 45, 9 (2021), 13489–13530.
- [9] Elena Giacomazzi, Felix Haag, and Konstantin Hopf. 2023. Short-term Electricity Load Forecasting using the Temporal Fusion Transformer: Effect of Grid Hierarchies and Data Sources. In *The 14th ACM International Conference on Future Energy Systems*. ACM, Orlando, FL, USA. <https://doi.org/10.1145/3575813.3597345>
- [10] Gene V. Glass. 1976. Primary, secondary, and meta-analysis of research. *Educational researcher* 5, 10 (1976), 3–8.
- [11] Marc Gürtler and Thomas Paulsen. 2018. Forecasting performance of time series models on electricity spot markets: a quasi-meta-analysis. *International journal of energy sector management* 12, 1 (2018), 103–129.
- [12] Stephen Haben, Siddharth Arora, Georgios Giasemidis, Marcus Voss, and Danica Vukadinović Greetham. 2021. Review of low voltage load forecasting: Methods, applications, and recommendations. *Applied Energy* 304 (2021), 1–37.
- [13] Mahmoud A. Hammad, Borut Jereb, Bojan Rosi, and Dejan Dragan. 2020. Methods and models for electric load forecasting: a comprehensive review. *Logistics, Supply Chain, Sustainability and Global Challenges* 11, 1 (2020), 51–76.
- [14] Julian Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew Page, and Vivian Welch. 2022. *Cochrane Handbook for Systematic Reviews of Interventions: Version 6.3*. www.training.cochrane.org/handbook
- [15] Tao Hong, Pierre Pinson, Yi Wang, Rafał Weron, Dazhi Yang, and Hamidreza Zareipour. 2020. Energy forecasting: A review and outlook. *IEEE Open Access Journal of Power and Energy* 7 (2020), 376–388.
- [16] Konstantin Hopf. 2018. Mining Volunteered Geographic Information for Predictive Energy Data Analytics. *Energy Informatics* 1:4 (2018). <https://doi.org/10.1186/s42162-018-0009-3>
- [17] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer Texts in Statistics, Vol. 103. Springer, New York, NY. <http://link.springer.com/10.1007/978-1-4614-7138-7>
- [18] Klaus Krippendorff. 2018. *Content analysis: an introduction to its methodology* (fourth edition ed.). SAGE, Los Angeles.
- [19] Corentin Kuster, Yacine Rezgui, and Monjur Mourshed. 2017. Electrical load forecasting models: A critical systematic review. *Sustainable Cities and Society* 35 (2017), 257–270.
- [20] Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37, 4 (Oct. 2021), 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>
- [21] Jon P. Nelson and Peter E. Kennedy. 2009. The use (and abuse) of meta-analysis in environmental and natural resource economics: an assessment. *Environmental and resource economics* 42, 3 (2009), 345–377.
- [22] Isaac Kofi Nti, Moses Teimeh, Owusu Nyarko-Boateng, and Adebayo Felix Adekoya. 2020. Electricity load forecasting: A systematic review. *Journal of Electrical Systems and Information Technology* 7, 1 (2020), 1–19.
- [23] Maamar Sebri. 2016. Forecasting urban water demand: A meta-regression analysis. *Journal of environmental management* 183 (2016), 777–785.
- [24] T. D. Stanley and Hristos Doucouliagos. 2012. *Meta-Regression Analysis in Economics and Business*. Taylor & Francis Group, Florence, UNITED KINGDOM. <http://ebookcentral.proquest.com/lib/ub-bamberg/detail.action?docID=1016122>
- [25] Tom D. Stanley, Hristos Doucouliagos, Margaret Giles, Jost H. Heckemeyer, Robert J. Johnston, Patrice Laroche, Jon P. Nelson, Martin Paldam, Jacques Poot,

- and Geoff Pugh. 2013. Meta-analysis of economics research reporting guidelines. *Journal of economic surveys* 27, 2 (2013), 390–394.
- [26] Ying Sun, Fariborz Haghighat, and Benjamin C. M. Fung. 2020. A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings* 221 (2020), 1–50.
- [27] Varun Varghese, Makoto Chikaraishi, and Junji Urata. 2020. Deep learning in transport studies: A meta-analysis on the prediction accuracy. *Journal of Big Data Analytics in Transportation* 2, 3 (2020), 199–220.
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., 5998–6008. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- [29] Eliana Vivas, Héctor Allende-Cid, and Rodrigo Salas. 2020. A systematic review of statistical and machine learning methods for electrical power forecasting with reported mape score. *Entropy* 22, 12 (2020), 1–24.
- [30] Jessica Walther and Matthias Weigold. 2021. A systematic review on predicting and forecasting the electrical energy consumption in the manufacturing industry. *Energies* 14, 4 (2021), 1–24.
- [31] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. 2018. Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid* 10, 3 (2018), 3125–3148.
- [32] H. White. 1980. A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica* 48 (1980), 817–838.
- [33] Achim Zeileis. 2004. Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software* 11, 1 (Nov. 2004), 1–17. <https://doi.org/10.18637/jss.v011.i10> Number: 1.