Haag, Felix; Hopf, Konstantin; Staake, Thorsten

Validating Explainer Methods : A Functionally Grounded Approach for Numerical Forecasting

Primary publication

Haag, Felix; Hopf, Konstantin; Staake, Thorsten (2026): Validating Explainer Methods : A Functionally Grounded Approach for Numerical Forecasting, in: Journal of Forecasting, New York, NY: Wiley Interscience, Vol. 45, Nr. 2, pp. 819–836, doi: 10.1002/for.70060.

Journal of Forecasting

WILEY

RESEARCH ARTICLE OPEN ACCESS

Validating Explainer Methods: A Functionally Grounded Approach for Numerical Forecasting

Felix Haag[1] | Konstantin Hopf[1,2] | Thorsten Staake[1,3]

[1]Chair of Information Systems and Energy Efficient Systems, University of Bamberg, Bamberg, Germany | [2]Chair of Information Systems and Business Analytics, Chemnitz University of Technology, Chemnitz, Germany | [3]Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland

Correspondence: Felix Haag (felix.haag@uni-bamberg.de)

Received: 4 April 2024 | **Revised:** 29 August 2024 | **Accepted:** 10 October 2025

Funding: The research presented in this paper was financially supported by the EUREKA member countries and the European Union (Eurostars grant number E!114466 - BENEFIZZO).

Keywords: explainable artificial intelligence | explainer method validation | explanation quality | interpretable machine learning | numerical forecasting

ABSTRACT

Forecasting systems have a long tradition in providing outputs accompanied by explanations. While the vast majority of such explanations relies on inherently interpretable linear statistical models, research has put forth eXplainable Artificial Intelligence (XAI) methods to improve the comprehensibility of nonlinear machine learning models. As explanations related to forecasts constitute important building blocks in forecasting systems, the validation of explainer methods is an essential part of system selection, parameterization, and adoption. Current research on explainer method assessment focuses on metrics for classification rather than numerical forecasting and predominantly assesses explanation quality within time-consuming, costly, and subjective studies involving humans. Given that the functional validation of explanations is of core interest to research on forecasting, our paper makes three contributions: First, we establish an approach for functionally grounded validations of explainer methods for numerical forecasting. Second, we propose computational rules for the metrics consistency, stability, and faithfulness. Third, we demonstrate our approach for the forecasting case of electricity demand estimation for energy benchmarks and compare a linear statistical approach with the state-of-the-art XAI methods SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Explainable Boosting Machine (EBM). Our work allows research and practice to validate and compare the quality of explainer methods on a functionally grounded level.

1 | Introduction

Decision-making is a process of constant information gathering (Aspers 2018). To facilitate this activity, companies employ forecasting systems to foster more informed business decisions (Doszyń 2019). Through recent developments in the area of machine learning (ML), such systems have made strong advances by employing ML applications, for example, in the area of sales (De Castro Moraes et al. 2024; Theising et al. 2023), finance (Liu et al. 2023; Zhang et al. 2024), and demand planning (Ducharme et al. 2024; Roach et al. 2021). As corporate decisions are usually made in complex environments, forecasts are only one source of information. In particular, explanations related to forecasts promise significant benefits for decision-makers, as they provide further insights on model outputs. Indeed, numerous scholars show that explanations alongside forecasts can support decision-making by improving decision performance (Bansal et al. 2021; Lai et al. 2020), trust (Yeomans et al. 2019), and the acceptance of machine-generated advice (Önkal et al. 2009). In response to the importance of explanations for decision-making processes, research has long been investigating how to effectively embed explanations in forecasting systems (Gregor and Benbasat 1999).

The vast majority of the established explainer methods rely on linear models (Fischer et al. 2016; Peters 2001). Due to their inherent interpretability, linear models are common in many areas of statistics and explanatory data analysis (Shmueli and Koppius 2011). Yet, linear models frequently suffer from low model fit, which can result in low predictive performance (Breiman 2001; James et al. 2013; Shmueli 2010). To overcome the limited predictive power of linear models, ML—one of the core technologies of artificial intelligence—finds increasing use within forecasting systems (Zhang et al. 2023). ML algorithms can estimate nonlinear relationships and are proven to work reliably with large amounts of business data (Chen et al. 2024; Jiang and Zheng 2022). However, resulting complex ML models, such as neural networks with numerous layers (known as deep learning methods), are composed of complex structures, making it difficult to explain why a model makes a certain decision. Complex ML models are therefore also referred to as "black boxes" (Adadi and Berrada 2018; Guidotti et al. 2018) and come along with issues regarding transparency, auditability, and accountability (Liu et al. 2024). Motivated by this accuracy-interpretability trade-off, research has put forth eXplainable Artificial Intelligence (XAI) techniques that translate patterns discovered by ML into human-readable form (Adadi and Berrada 2018; Barredo Arrieta et al. 2020; Bauer et al. 2021). Although research on the explainability of forecasts dates back to the 1980s (Swartout 1983), the increasing complexity of recent ML models recently made XAI a very active field of research (Barredo Arrieta et al. 2020; Liu et al. 2023). In addition to XAI techniques that explain complex (black-box) ML models post hoc (Lundberg et al. 2020), recent research has put forth models that come with comparably high predictive power but also possess capabilities to explain their outputs (Kraus et al. 2023). Consequently, forecasting research has applied XAI methods in various cases employing multiple categories of methods (Lim et al. 2021; Liu et al. 2023; Wu et al. 2024).

So far, many studies that apply XAI methods *evaluate* the explanations generated, for example, in terms of user perception, trust, and acceptance (Gregor and Benbasat 1999; Shin 2021). Beyond these rather cost-intensive and subjective application- and user-oriented evaluations, the *validation* of information systems and their components (e.g., a forecasting engine) is necessary for "checking of the appropriateness of the system for the purpose for which it is being used." (Finlay 1994, 209). In the same vein, XAI research employs three levels of explainer method assessment: (i) human- and (ii) application-grounded evaluations (both involving human subjects), and (iii) functionally grounded validations, whereas validation equates to an assessment using proxy tasks as a precondition for modeling real-world relationships in data (Doshi-Velez and Kim 2017). Functional-grounded validation is a fundamental step in the scientific and effective development of computerized systems. The aim is to prove that the forecasting model has identified suitable underlying relationships that reflect the real world in a particular domain (Borenstein 1998; O'Leary 1987). From a practical perspective, functional validation is repeatable without the need to recruit potentially biased study participants. It can also be performed with uncritical test data sets, as may be required, for example, in the selection of explainer methods (Doshi-Velez and Kim 2017).

Current research on explainer method validations focuses on metrics for classification tasks rather than numerical forecasting tasks (e.g., Velmurugan et al. 2021a; Schlegel et al. 2019) and either on XAI (e.g., Alvarez Melis and Jaakkola 2018; Velmurugan et al. 2021b) or inherently interpretable linear statistical method validation (e.g., Mészáros and Rapcsák 1996). Given that explanations alongside numeric forecasts are of core interest for research on forecasting systems, our paper makes three main contributions: First, we establish an approach for functionally grounded validations and comparisons of explainer methods for numerical forecasting. Second, we use the already conceptually defined metrics *model fit*, *consistency*, *stability*, and *faithfulness* (Robnik-Šikonja and Bohanec 2018) and propose concrete computational rules for these metrics focusing on numerical forecasting—so far, such rules are only known for classification tasks (e.g., Schlegel et al. 2019; Velmurugan et al. 2021b). Third, our approach allows for the validation and comparison of explainer methods across various categories of methodological approaches (e.g., linear statistical, XAI, and white-box explainer methods). We demonstrate our approach using a forecasting case in which several categories of explainer methods are common to derive influencing factors, namely energy benchmarking (Arjunan et al. 2020; Huebner et al. 2015). In this demonstration, we compare a linear statistical approach using multiple linear regression (MLR) with the state-of-the-art XAI methods SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and the white-box approach Explainable Boosting Machine (EBM). Our work allows research and practice to validate and compare the goodness of explainer methods within forecasting systems on a functionally grounded level beyond rather costly and subjective evaluations involving human subjects. More specifically, the results of our work provide researchers and system developers with an instrument with which they can validate and qualify explanations in forecasts for practical use. Such an instrument is essential for a meaningful integration of forecasting systems into a decision-making environment.

2 | Background

Our approach draws on literature regarding forecasting systems, XAI, and explainer methods as well as their evaluation and validation. In the following sections, we review the literature of the related research domains.

2.1 | Characteristics of Explainer Methods

Given the wide range of explainer methods (Lundberg and Lee 2017; Moreira et al. 2021; Ribeiro et al. 2016), literature has put forth several characteristics to categorize XAI techniques (Adadi and Berrada 2018; Barredo Arrieta et al. 2020; Meske et al. 2020). We summarize four of such characteristics that are relevant for our work. The first characteristic differentiates between *post hoc* and *intrinsic interpretable* methods (Rai 2020). It describes whether a forecasting model needs any additional procedures to enable human interpretability. Methods that explain an already trained model, for example, by permuting feature values and presenting their influence on the forecast in a human-readable way, are referred to as post hoc methods. In contrast,

intrinsic approaches are considered interpretable by design (e.g., linear regression) due to their simplicity and because they do not require any further methods to increase the tractability of model decisions. The second characteristic distinguishes between *model-agnostic* and *model-specific* explainers. Model-agnostic methods are applicable to any model, which makes them independent of specific model architectures or types. Model-specific approaches, as the name suggests, are limited to a certain class of algorithmic approaches (e.g., tree-based ML). Intrinsically interpretable methods are considered inherently model-specific. A third characteristic refers to the degree of interpretability: *Local interpretable* methods have the capability to explain forecasts on an individual level (e.g., influence of a feature value on the predicted outcome). Thus, they allow for elaborating why a model has come to an individual decision in a particular case. *Global interpretable* explainer methods focus on the model as a whole and try to explain the overall logic and behavior that lead to all outcomes. A fourth characteristic considers the type of explainer for generating and presenting model explanations (Adadi and Berrada 2018): *Example-based* methods attempt to explain the model behavior using a few selected observations of the dataset. Explanations of *knowledge extraction* methods focus on the rules of complex models (e.g., artificial neural networks) and try to visualize knowledge that models acquire during training. The concept of influence methods attempts to estimate the importance or attribution of individual features by modifying data or model components and observing changes in the model outcome. In particular, *feature attribution* methods, which are model-agnostic and often allow for both the estimation of local and global explanations, recently gained prominence in the field of influence methods and XAI in general. Two frequently cited approaches that match these criteria are SHAP and LIME, which we describe in more detail in Section 4.3.

2.2 | Evaluation and Validation of Explainer Methods

The design and development of new functionalities in forecasting systems involves a thorough testing against system requirements, the current state-of-the-art (Phillips-Wren et al. 2009), and real-world validity (Borenstein 1998). For quality tests of expert systems such as forecasting systems (O'Leary 1987), literature has pointed to differences between evaluation and validation. *Evaluation* is the process of assessing the overall benefit of a software system "to the users, project sponsors and ultimately the organization, and is generally associated with measures of worth and value for money" (O'Keefe and O'Leary 1993, 5). *Validation*, by contrast, describes "the process of defining whether the model behavior represents the real world system in a particular problem domain" (Borenstein 1998, 227).

Focusing on explainer method *evaluation*, the literature recognizes two approaches (Doshi-Velez and Kim 2017): First, *application-grounded* evaluations, relying on specific tasks in a particular application context and feedback from human experts in the field. Here, literature suggest metrics such as *benefits of the explanation* for a specific use case, *persuasiveness, completeness*, and *novelty* of information provided to the application field to measure how humans cope with the explanations produced in a specific field of application (O'Keefe and O'Leary 1993; Schwalbe and Finzel 2023). The second approach is *human-grounded* evaluations, which assess data-driven explanations by providing individuals with simplified tasks (e.g., laypeople or specific samples) in controlled lab experiments. Examples are the comparative evaluation of several explanation approaches or visualizations in survey-based experiments (Lakkaraju et al. 2016; Wastensteiner et al. 2021). For human-grounded evaluations, the degree of *understanding*, *interpretability*, and *effectiveness* serve as measures to evaluate the extent to which an explanation receiver can make sense and build a mental model of the explanations provided (Schwalbe and Finzel 2023). Both application-grounded and human-grounded evaluations (i.e., human-subject experiments) often follow paradigms from human-computer interaction research (Abdul et al. 2018; Antunes et al. 2008; Shin 2021) to examine a "system's ability to solve real-world problems in a particular problem domain" (O'Leary et al. 1990, 51). This research stream aims to increase the understanding of how humans make sense of visualizations (Lee et al. 2016) and derive recommendations for visualizing explanations and eliciting user preferences (Hudon et al. 2021). Examples of such evaluations that take human judgment into account range from applications in healthcare (Branley-Bell et al. 2020) to the energy sector (Wastensteiner et al. 2021). The inclusion of human preferences, however, places an additional layer of complexity on the evaluation of explanations, which makes systematic analysis and comparison of explainer methods difficult. In addition, these approaches require implementation of the explainer in real-world applications or lab experiments, which is a time-consuming and expensive effort.

Explainer method validation studies, conversely, focus on the functionality by employing scientific assessments of computer-based systems that are more objective (O'Keefe and

TABLE 1 | Overview of explanation quality assessment approaches and their corresponding objectives and metrics.

Approach	Application-grounded	Human-grounded	Functionally grounded
Objective	*Evaluation* of explanations for a specific real-world application	*Evaluation* of general notions of explanation quality	*Validation* of explanations using proxy tasks
Metrics	• Benefits of the explanation (for a specific use case) • Persuasiveness • Completeness • Information novelty for application field • ...	• Understanding • Interpretability • Effectiveness • ...	• Model fit • Consistency • Stability • Faithfulness

O'Leary 1993; O'Leary et al. 1990). Similarly, Doshi-Velez and Kim (2017) call for functionally grounded approaches to validate XAI methods, using metrics that do not involve humans and address methodological properties. Metrics that belong to this category are, for example, *model fit*, *consistency*, *stability*, and *faithfulness*[1] (Robnik-Šikonja and Bohanec 2018) (Table 1).

Yet, literature has so far defined functionally grounded metrics widely only on a conceptual level, and recent studies that consider such metrics for validating and comparing explainer methods focus on classification rather than numerical forecasting tasks (Alvarez Melis and Jaakkola 2018; Schlegel et al. 2019). To date, literature has mostly established clear computational rules for the *model fit* in numerical forecasting, but not for the metrics' *faithfulness*, *consistency*, and *stability*. Therefore, one of the objectives of our work is to formally define meaningful computational rules for these metrics. Without the standardized calculation of such metrics, "it is neither clear for all these properties how to measure them correctly nor how useful they are to specific use cases, so one of the challenges is to formalize how they could be calculated" (Carvalho et al. 2019, 16–17).

3 | FEXVAL: Functionally Grounded EXplanation VALidation Approach for Numerical Forecasting

In response to the lack of clearly described procedures for the functionally grounded validation of explanations, we establish FEXVAL as an approach that allows for the validation of quantitative feature-attribution explainer methods for numerical forecasting. Our approach is suitable to validate and compare two or more explainer methods and helps forecasting system developers select a suitable method without costly evaluations involving human subjects. For this purpose, FEXVAL uses four metrics that are defined by current literature and do belong to the category of functionally grounded validations (Table 1).

Figure 1 outlines the procedure for validating and comparing explainer methods in five steps. The first step involves a method-independent data preprocessing, which includes the partitioning of the data into a training and a test set. The subsequent Steps 2–4 are then performed individually for each method under consideration. In the second step, the forecasting model undergoes training and then an assessment using the test data and the metric *model fit (i)*. Although this step assesses the model's forecasting quality rather than the explanations themselves, it serves as an initial measure for explanation validation. This procedure is grounded in prior research, which suggests that models with high predictive accuracy are likely to produce higher-quality explanations (Štrumbelj et al. 2009). The third step focuses on generating explanations for the forecasting model. For black-box models, this requires applying a corresponding post hoc method, such as LIME or SHAP. White-box models, on the other hand, do not require this additional procedure due to their inherent interpretability. To ensure the comparability of explanations across different methods, it may be necessary to adjust the output accordingly (see, e.g., Section 4.3.3). In the fourth step, the explanatory output is further validated using the metrics of *consistency (ii)*, *stability (iii)*, and *faithfulness (iv)*. When an explainer method performs well in one metric but poorly in another, FEXVAL provides an optional approach for calculating a weighted trade-off. This allows for the selection of an appropriate explainer method tailored to the specific needs of the forecasting system in step five. We describe the functionally grounded metrics and trade-off calculation in more detail below.

3.1 | Model fit

In general, we can accept a model that is capable of explaining its outputs (i.e., through inherent interpretability or a post hoc explainer method) but has a low *model fit* and thus low predictive performance (Molnar 2019; Shmueli 2010). Empirical studies, however, demonstrate that forecasting models with a higher predictive performance also allow for better explanations (Štrumbelj et al. 2009). This is due to the possibility that explanations based on predictions of an under-fitted or over-fitted model might lead to spurious (real-world) conclusions. We therefore argue that *model fit* is a necessary but not sufficient criterion for judging the quality of explainer methods related to forecasts.

We consider *model fit* to describe how well an (underlying) explanation prediction method can predict and thus can generalize on yet unseen observations (Robnik-Šikonja and Bohanec 2018; Štrumbelj et al. 2009). To measure *model fit*, we propose relying on established and well-known numerical forecasting quality measures (Table 2). Given that all measures show weaknesses in various aspects (see, e.g., Armstrong and Collopy 1992; Kim and Kim 2016), we suggest the following, depending on the use case: R^2 serves for a baseline assessment of model fit and overall main criterion, as it is a measure that describes the proportion of the explained variance in the target variable and is thus an indicator for model fit. If isolated substantial deviations from the actual value are costly, the RMSE can serve as a suitable criterion. If overall high performance is important (i.e., isolated substantial deviations are negligible), MAE and MAPE are more appropriate metrics to focus on. For more details on *model fit* and numerical forecasting error metrics, we refer to Hastie et al. (2009).

3.2 | Consistency

In the literature, the metric *consistency* has already been conceptually described as the extent to which different forecasting models lead to similar explanations when given the same data (Robnik-Šikonja and Bohanec 2018). While the forecasts of models trained on the same data may be similar, the explanations may vary. This can be due to a different weighting of the features by the model or a different way in which the explainer method obtains explanations. Beneath *consistency* lies the assumption that consistent explainer methods produce more robust and meaningful results (Molnar 2019). However, measuring consistency between explanations of various methods and models can be difficult due to the different features used by the (underlying) models (Molnar 2019).

We propose to assess *consistency* for a single explainer method by using a measure of reliability akin to *stability*, namely *internal consistency*. Employing *internal consistency* (hereafter also referred to as *consistency*) allows for assessing if an

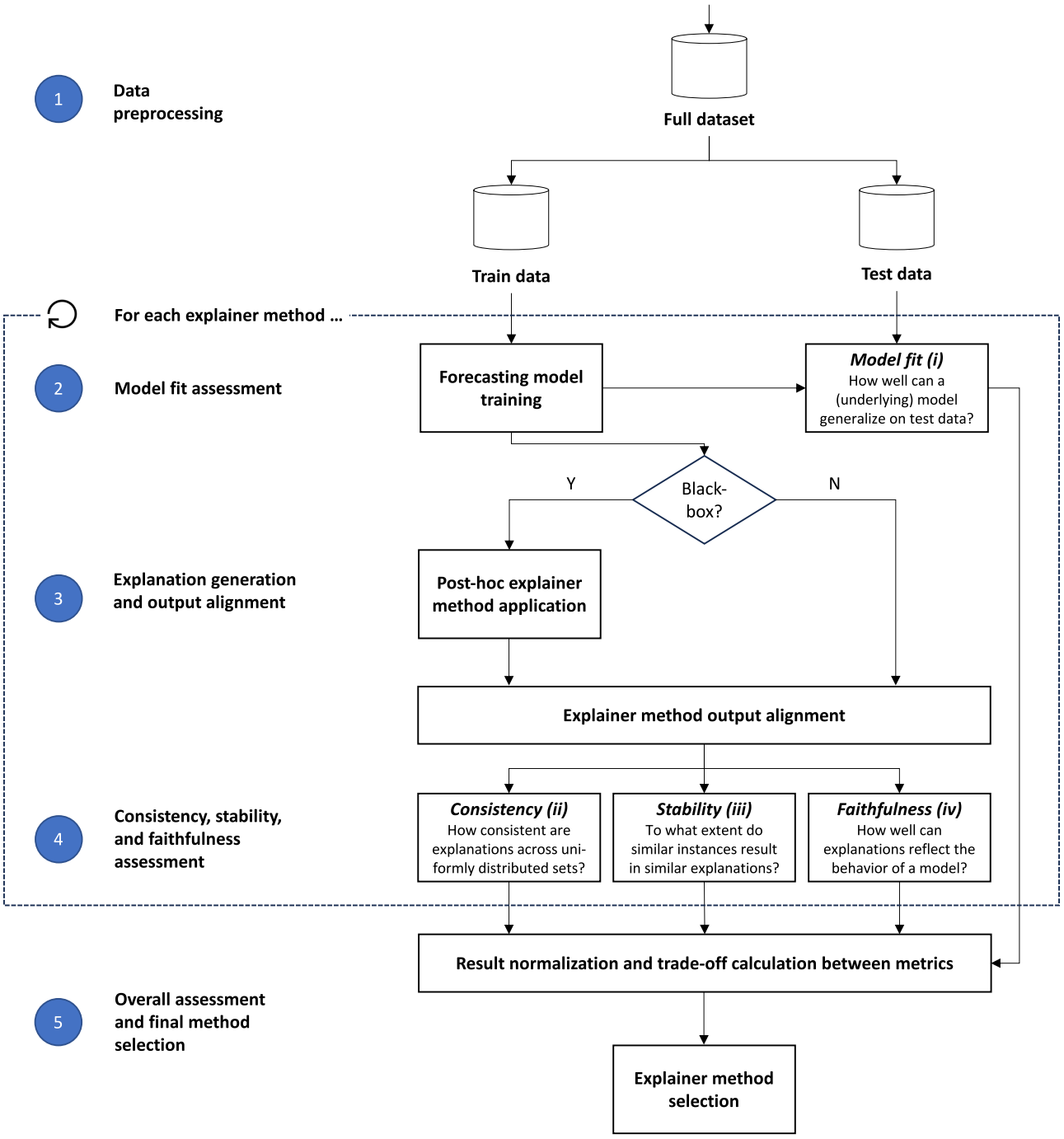


FIGURE 1 | Illustration of the FEXVAL approach to validate and compare explainer methods.

explainer method yields similar feature attributions in various subsamples comprising observations that are uniformly distributed according to the target variable. In other words, the analysis relies on the assumption that repeatedly drawing subsamples from an entire sample that are equally distributed with respect to their target variable should lead to consistent explanations because the observations are to some extent inherently similar. We instantiate the measure by considering k stratified folds and examining whether the explanations are consistent between folds. We define a method to be internally consistent for the respective feature if it assigns similar feature attributions that lead to a nonsignificant difference (i.e., homogeneous or consistent variances) in attributions between folds. We base consistency on a nonparametric Fligner-Killeen test (Fligner and Killeen 1976). We use the test to assess the difference of variances between folds, as the resulting explanation values (i.e., feature attributions) might be nonnormally distributed. The Fligner–Killeen test calculates as follows for the feature j:

$$\chi_j^2 = \frac{\sum_{i=1}^{k} n_i \left(\overline{A}_i - \overline{a}\right)^2}{V}, \tag{5}$$

where n_i denotes the size of the ith fold, $\overline{A}_i$ the arithmetic mean of normalized ranked values of the fold i, $\overline{a}$ the arithmetic mean of normalized values across all folds k, and V the variance of

TABLE 2 | Model fit measures.

Error metric	Description	Formula[a]
R^2	Relative proportion of the variance in the target variable that the predictor variables explain.	$\frac{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}$ (1)
Mean absolute error (MAE)	Average absolute deviation of the predicted from the actual value.	$\frac{1}{N}\sum_{i=1}^{N}\lvert\hat{y}_i - y_i\rvert$ (2)
Mean absolute percentage error (MAPE)	Average absolute percentage deviation of the predicted value from the actual value.	$\frac{100\,\%}{N}\sum_{i=1}^{N}\left\lvert\frac{\hat{y}_i - y_i}{y_i}\right\rvert$ (3)
Root mean squared error (RMSE)	Standard deviation of unexplained variance (spread of the residuals).	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2}$ (4)

[a] y_i = actual target value; $\hat{y}_i$ = model prediction; $\bar{y}$ = mean of actual target values; N = number of observations.

all normalized values (Conover et al. 1981). Based on this value, we determine the p value per feature and rate the consistency (i.e., homogeneity) of the variances between the folds based on a specified significance level α. Finally, we determine the *internal consistency* of a method by the percentage of consistent features across all features (i.e., the proportion of features for which we find a nonsignificant difference across folds).

3.3 | Stability

The metric *stability* describes the extent to which explanations are coherent with similar observations, that is, similar observations lead to similar explanations (Alvarez Melis and Jaakkola 2018). High *stability* means that the method can reliably explain learned patterns and relationships discovered by a model. Low *stability*, in turn, results from divergent explanations for observations that are alike. The literature suggests several methods for calculating stability for classification tasks. One approach measures whether a method selects similar features as being equally important (i.e., by examining the variability of feature weights) for instances that are assumed to be similar (Mohana Chelvan and Perumal 2016; Velmurugan et al. 2021b). Here, similarity can be defined, for example, by the membership in a certain class (Wastensteiner et al. 2021) or by grouping through instance-specific metadata (Velmurugan et al. 2021b). Another approach relies on the variation of observations by adding noise. The generation of stable explanations is assumed if the explanation changes only slightly, given that the varied input instances still rely on similar feature values (Alvarez Melis and Jaakkola 2018; Molnar 2019). To the best of our knowledge, no computational rule for stability in the context of numerical forecasts is known in the body of present literature.

We propose a stability validation procedure suitable for numerical forecasting, which allows for the consideration of datasets that contain numerical and categorical data. Our procedure comprises three steps to evaluate explainer method's *stability*. First, determining dissimilarities between observations using Gower's (1971) distance, which is applicable to data containing numerical and categorical features. For this purpose, Gower's distance expresses the average partial dissimilarities across all observations. Second, we suggest applying clustering, preferably using the Partitioning Around Medoids (PAM) algorithm that selects k actual data points as cluster centers (so-called medoids) to obtain subsets of similar instances (Kaufman and Rousseeuw 1990). Compared with k-means, k-medoid-based approaches do not sum squared distances but minimize the dissimilarities between observations pairwise, which makes the resulting clusters more robust to outliers and noise in the data. Third, we measure *stability*—adapted from Nogueira et al. (2018)—by defining h_j, which indicates the number of clusters k (i.e., subsets of similar observations) where feature j has the highest impact on the models' prediction according to an explainer method. Given p as the total number of available features and $q = \sum_{j=1}^{p} h_j$, then

$$Stability = 1 - \frac{\frac{1}{p}\sum_{j=1}^{p}\frac{k}{k-1}\frac{h_j}{k}\left(1-\frac{h_j}{k}\right)}{\frac{q}{kp}\left(1-\frac{q}{kp}\right)}, \tag{6}$$

measures *stability* for explainer methods ranging from 0 to 1, whereas 0 indicates no overlap between chosen features and 1 that all subsets are equal (i.e., maximum *stability*). Several cluster sizes should be used to obtain a holistic assessment of *stability*.

3.4 | Faithfulness

The metric *faithfulness*—also referred to as "fidelity" (Robnik-Šikonja and Bohanec 2018)—here quantifies to what extent features identified as relevant by explainer methods are truly relevant (Robnik-Šikonja and Bohanec 2018; Alvarez Melis and Jaakkola 2018). Thereby, it assesses whether explainer methods provide meaningful explanations for a given model output. Quantifying *faithfulness* requires knowledge of a feature's true impact. As the true impact is usually unknown for real-world datasets, literature suggests testing *faithfulness* with "blurring" input data, also called "perturbing" (Alvarez Melis and Jaakkola 2018; Du et al. 2019). In doing so, the features that explainer methods consider as most relevant are modified in a way that the model output is expected to change. Blurring of features can be done, for example, by replacing feature values with zeros

(Schlegel et al. 2019), the mean, or setting coefficients of additive models to zero (Alvarez Melis and Jaakkola 2018). Such an operation can be expected to change the prediction in the opposite direction for classification tasks. Following this assumption, existing literature suggests measuring the relative amount of prediction changes to the opposite class (Schlegel et al. 2019; Wastensteiner et al. 2021), or the average percentage decrease in the probability of a predicted class after noise has been added to the input (Du et al. 2019).

In our literature search, we could not find a computational approach to measure *faithfulness* for numerical forecasting. Drawing on the approach of Velmurugan et al. (2021a) for classification, we propose to compute *faithfulness* for numerical forecasting by using the average change in the model's output (expressed here as percentage change):

$$Faithfulness = \frac{\sum_{i=1}^{N} \frac{\left|\hat{f}(x_i) - \hat{f}(x_i')\right|}{\hat{f}(x_i)}}{N}, \tag{7}$$

where for a sample of N observations $x_1 \ldots x_n$, $\hat{f}(x_i$ describes the corresponding initial model output. For each x_i exists a corresponding perturbed observation x_i', whose values have been replaced with, for example, zero or the feature's mean value. $\hat{f}(x_i'$ denotes the model's output with perturbed feature values.

Given that we observe a numerical output, we suggest examining *faithfulness* from two perspectives using the mean absolute change in prediction. First, the perturbation of each feature value in the test dataset that is considered most relevant, respectively, to increase the prediction. Second, the modification of the feature values in the test set that are most relevant for lowering the model output according to the respective explainer method. If we observe a considerable shift of the prediction in the opposite direction, we assume that a method produces *faithful* explanations.

3.5 | Trade-Off Between the Metrics

Although the four metrics we introduce—model fit, consistency, stability, and faithfulness—are independent and assess different aspects of an explainer method, there may be scenarios where it is necessary to calculate a trade-off between these metrics. This is particularly important when a single performance measure is required for each explainer method. To address this need, we recommend computing an overall score using the geometric mean of all metrics. The geometric mean offers an approach to aggregation of normalized values (Fleming and Wallace 1986) and is in this case preferable over both the harmonic and arithmetic means as it fairly aggregates performance across all metrics while still effectively penalizing low performance in any single metric.

Before calculating the trade-off and deciding for a method, it is crucial to normalize the results to ensure comparablity (Corrente and Tasiou 2023). We suggest applying min-max normalization to scale the metric values between 0 and 1. However, min-max normalization can result in zero for some metrics, which could lead to an unfair overall score of zero when using the geometric mean. To prevent this, we recommend adjusting each metric result M_{mr} for the explainer method m and metric r by adding a small positive constant η (e.g., $\eta = 0.01$) to any metric that is zero:

$$M'_{mr} = \begin{cases} M_{mr} + \eta & if\ M_r = 0 \\ M_{mr} & if\ M_r > 0 \end{cases}, \tag{8}$$

where $r \in \{1, \ldots, 4\}$ describes the four metrics. Here, $r = 1$ denotes *model fit*, $r = 2$ *consistency*, $r = 3$ *stability*, and $r = 4$ *faithfulness*. This adjustment ensures that the geometric mean can be calculated without the risk of it being entirely nullified by a zero in any metric result.

Once the metrics have been normalized and adjusted, the performance $P(m)$ of an explainer method m can be evaluated using the geometric mean of the metrics. When all metrics are considered equally important, the performance is computed as follows:

$$P(m) = \sqrt[4]{\prod_{r=1}^{4} M'_{mr}}, \tag{9}$$

where each metric is considered equally important. However, in some cases, it may be suboptimal to assume that all metrics are equally important (e.g., when *stability* is the main criterion). For scenarios where certain metrics are preferred over others, a weighted geometric mean can be employed. The overall performance of each method is then computed with

$$P^*(m) = \prod_{r=1}^{4} \left(M'_{mr}\right)^{\alpha_r}, \tag{10}$$

where $\alpha_r \in (0, 1)$ are the weights assigned to each metric r, with $\sum_r \alpha_r = 1$. The weights for the metrics must be obtained in relation to the case and specific needs of a forecasting system. One approach would be, for example, to use pairwise comparisons using the fundamental scale, as known from the analytical hierarchy process (Saaty 2008) or other approaches known from managerial decision making (Goodwin and Wright 2014).

4 | Demonstration of FEXVAL Using Energy Benchmarking as a Use Case

To demonstrate FEXVAL, we choose residential energy benchmarking for electricity consumption as a use case for three reasons: First, the energy demand estimation, as frequently required for energy benchmarking, is a common use case for forecasting systems (Fan et al. 2020; Roach et al. 2021). Second, linear statistical models have a long tradition in explaining the influencing factors (i.e., features) of energy consumption (Huebner et al. 2015; Santin et al. 2009), but XAI methods have already been tested for energy benchmarking (Arjunan et al. 2020). Third, the data for energy benchmarking contain patterns that can be causally related to user characteristics and behavior (e.g., number of residents and frequency of load peaks).

Energy benchmarking uses a reference point (e.g., a score) to compare energy performance with a suitable group that shares similar characteristics, usually in terms of specific household or building properties (Arjunan et al. 2020; Palmer and Walls 2015). The approach is a useful tool to identify inefficient buildings, appliances, and behavior, thus helping stakeholders to manage energy consumption more efficiently (Chung 2011). The concept of benchmarking becomes even more important when focusing on electricity, as the electricity demand is expected to double by 2050 through the increased usage for transportation, heating, and sector coupling (European Commission 2012). However, energy benchmarks are often difficult to interpret, as they are usually limited to a predicted single score—in other words, energy benchmark users have difficulty identifying the factors that affect the performance of their housing unit (Arjunan et al. 2020). To enhance energy benchmarks with targeted insights, several works use statistical methods such as MLR to narrow down factors that dominate energy consumption (Huebner et al. 2015). Considering the limitations of linear statistical models, comparing them with XAI methods is a reasonable effort to improve the insights obtained from electricity benchmarking.

4.1 | Benchmarking Methodology

Given that households differ in terms of physical (e.g., living space, glazing, and heating) and resident characteristics (e.g., number of adults and children), energy benchmarking relies on performance indicators, as the heterogeneity of characteristics can have a complex impact on energy consumption. Considering the high influence of such factors on the energy demand, energy benchmarking approaches usually normalize consumption. While there are several normalization approaches, the Energy Use Intensity (EUI) is one of the most common and simplest key metrics for comparison and has been often applied in research related to energy benchmarking (Arjunan et al. 2020; Chung 2011; Wang 2019). The EUI normalizes a building's yearly energy consumption over differences in floor area. In regions using metric units of measurement, the EUI computes as kilowatt-hours (kWh) per year/square meters (m^2).

4.2 | Data

For our quantitative analyses, we use a dataset from the Commission for Energy Regulation in Ireland (Commission for Energy Regulation 2011), which is publicly available on request. The dataset consists of electricity consumption data at 30-min intervals from households across Ireland between July 14, 2009, and December 31, 2010. In addition, data from two surveys (pre- and post-trial) contain characteristics of the participating households, from which we use 14 household characteristics (see Table 3).

Our data preparation and cleansing process comprises two steps. First, we removed households from the dataset that have missing values in the survey or the electricity consumption data. Second, we filtered for outliers, as these can have a strong influence on the results of the methods used. According to the Central Statistics Office Ireland (2010), the average floor area

TABLE 3 | Overview of household characteristics used.

Variable	Description
ageBuilding	Age of the building derived from the year in which the building was built.
dwellingType	Dwelling type considering apartment, detached house, semi-detached house, terraced house, and bungalow.
electricCooking	Whether a household cook electrically (yes = 1; no = 0).
electricShower	Whether a household uses instant electric or electric pumped from hot tank showers (yes = 1; no = 0).
electricSpaceHeating	Whether a household uses electric space heating (yes = 1; no = 0).
electricWaterHeating	Whether a household uses electric water heating (yes = 1; no = 0).
floorArea	Building's gross floor area (in m^2).
numberAdults	Number of persons over 15 years of age.
numberEntertainmentAppliances	Number of entertainment devices, including TVs, desktop computers, laptops, game consoles.
numberHomeAppliances	Number of domestic appliances, including washing machines, tumble dryers, dishwashers, plug-in convector heaters, freezers, water pumps/electric well pump, immersions/kettles.
numberKids	Number of persons under 15 years of age.
numberPeaks	Annual frequency of load peaks in electricity consumption.
proportionDoubleGlazed	Relative proportion of double-glazed windows in building (0–100%).
proportionEnergySavingBulbs	Relative proportion of energy-saving bulbs in building (0–100%).

of newly granted dwellings in 2010 was 84.5 m^2 and for already built houses 104 m^2 (Dol and Haffner 2010). Thus, we only include buildings that have a minimum floor area of 10 m^2 for apartments and 30 m^2 for houses (including bungalows) and are within five times the respective dwelling type's average. After these data cleaning steps, our final dataset contains 1262 observations.

4.3 | Explainer Method Implementation

We select four explainer methods that are currently common in the field. These methods belong to two explainer method categories and yield feature attributions: the post hoc explainer methods LIME and SHAP and the white-box approaches EBM and statistical MLR.

4.3.1 | LIME

LIME is a post hoc explainer method and obtains local explanations of any black-box prediction model (i.e., a model agnostic approach) (Ribeiro et al. 2016). LIME uses an optimization function $\mathcal{L}$, which searches for a local approximation to a complex model f for a given input x in the proximity π_x of this focal point. The local approximation is called surrogate model g from a family of sparse linear models G. Given that LIME searches for a simple explanation, the second loss term Ω regularizes the complexity of the simple surrogate model:

$$\xi(x) = argmin_{g \in G} \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g). \tag{11}$$

To obtain explanations, LIME randomly generates new instances in the neighborhood of a focal point x through perturbation, and $\mathcal{L}$ minimizes the prediction error of the surrogate model to explain the predictions of the complex model f. The sparse linear model used to obtain the explanations estimates coefficients for each feature, which are the feature-attributions of LIME. Although popular, LIME is a has recently been criticized for yielding nonreproducible results, as repeated explanations for the same or similar observations can lead to different explanations (Carvalho et al. 2019). Indeed, due to the random generation of artificial input instances, explanation results can vary between runs (Schlegel et al. 2019; Velmurugan et al. 2021b).

4.3.2 | SHAP

SHAP is an XAI method that uses a concept of game theory, namely, Shapley values (Lundberg and Lee 2017). Thereby, it considers every feature to be a player contributing to a common game, which is the forecast. Explanations provided by SHAP are local and represented by an inherently interpretable model. SHAP provides a whole class of additive feature importance measures that one can use to explain post hoc the output of any ML. In addition to the model-agnostic KernelSHAP implementation (Lundberg and Lee 2017), TreeSHAP is a model-specific implementation for tree-based ML algorithms such as Random Forest (RF) (Lundberg et al. 2020). SHAP defines the explanation model through a linear function of binary features:

$$g\left(v'\right) = \phi_0 + \sum_{j=1}^{F} \phi_j v_j, \tag{12}$$

where g represents the explainer model and $v' \in \{0,1\}^J$ are so-called "simplified features" describing the presence or absence of a feature as a vector, where J is the total number of features. F denotes the number of input features (i.e., the maximum size of a coalition of features) and $\phi_j \in \mathbb{R}$ belongs to the attribution (i.e., the SHAP value) of feature j, while ϕ_0 describes the expected value $E\left[\hat{f}(X)\right]$ across all observations (so-called "base value"). Summing up the base-value ϕ_0 and the marginal feature contributions ϕ_j yields the model's prediction $\hat{f}(x)$––hence, SHAP values explain the marginal attribution of features on the prediction as the deviation from a model's average outcome (Lundberg and Lee 2017; Molnar 2019). Several studies showed that SHAP outperforms LIME with respect to stability (Schlegel et al. 2019; Velmurugan et al. 2021b) and faithfulness (Velmurugan et al. 2021a).

4.3.3 | MLR

For the linear and intrinsically interpretable approach (i.e., a white box explainer method), we rely on standard ordinary least squares (OLS) weight estimation for our MLR model

$$Y_i = \beta_0 + \sum_{j=1}^{F} \beta_j X_{ij} + \epsilon_i, \tag{13}$$

where for each observation $i \in \{1, \ldots, n\}$, there is a target value Y_i. Here, X_i denotes a vector of F predictor variables (i.e., features), β_0 is the offset term, β_j the regression coefficient for feature j and ϵ_i describes the error for the ith observation (Hastie et al. 2009). MLR belong to the class of model-specific approaches that are designed to furnish users with global explanations. The linearity and additivity of MLR make it an intrinsically interpretable explainer method. The linearity property, on the one hand, refers to the constant weights for each explanatory variable. While linearity is desirable in the context of interpretability, they may not hold for real-world applications, which is why MLR models often found to have only low modeling capabilities (Arjunan et al. 2020). The additivity property, on the other hand, enables the effects to be separated, that is, the influence of a predictor on the model output is independent of all other variables. Therefore, both properties make it possible to understand the isolated impact of an explanatory variable on a specific outcome. To compute local feature attributions, we use a computational approach presented by Štrumbelj and Kononenko (2014) to determine a feature value's impact––after conversion, the interpretation of feature attributions is equal to those of SHAP values.

4.3.4 | EBM

EBM belongs to the class of generalized additive model (GAM) approaches that allow for global and local interpretations on a model-specific basis. GAMs independently map the input features in a nonlinear fashion and sum up these mappings using an additive link function g that adapts to numerical forecasting or classification tasks:

$$g(E[y]) = \beta_0 + \sum_{j=1}^{F} f_j(x_j), \quad (14)$$

where $f_j(x_j$ represents an individual shape function for each feature $j \in F$ that allows for the interpretation of how a feature affects the predicted output $E[y$. The interpretation of feature effects is thereby similar to that of an MLR model, as the contribution of features can be separated (i.e., additivity property). EBMs come with two advantages over conventional GAMs and MLRs. First, EBMs learn the shape function using modern bagging and boosting approaches while mitigating the negative effects that result from the collinearity of features. Second, they are capable of automatically detecting pairwise feature interactions (Lou et al. 2013). Yet, for comparability reasons with the other methods in our study, we do not include such interactions in the model. While EBM belongs to white-box approaches, it has already demonstrated in many cases to achieve comparably high model fit when compared with black-box ML (Lou et al. 2012; Nori et al. 2019; Zschech et al. 2022).

4.4 | Case Analysis and Findings

We exemplify the application of FEXVAL below and describe the processing steps necessary to validate and compare the selected explainer methods.

4.4.1 | Model Fit

For determining how well a (underlying) forecasting model can describe the relationships in the data on unseen data (i.e., the model fit), we use a split in training and test data to avoid overfitting, employ k-fold cross-validation to train models, and to find the optimal parameter configuration (Hastie et al. 2009). In addition, to avoid bias in error estimation due to an overfit on the train data, we perform the final validation on the remaining test data (i.e., the holdout set) (Rao et al. 2008). In doing so, we use a stratified 80/20 split in training and test data. We apply a cross-validated grid search with 10 folds for the ML models to determine the best parameter configuration. In addition, we use z-score normalization for numerical features to avoid algorithmic bias due to different scales of features and use one-hot encoding to handle categorical features.

To compare the model fit of the (underlying) models, we use MLR and EBM as the two inherently interpretable (i.e., white-box) approaches and draw on state-of-the-art ML approaches for the post hoc explainer methods. We selected ML methods from different classes (ensemble, boosting, neural net, and kernel-based algorithms) that recent studies have already applied to predict energy consumption (Arjunan et al. 2020; Robinson et al. 2017). These are eXtreme Gradient Boosting (XGBoost), CatBoost, and Light Gradient-Boosting Machine (LightGBM), which are based on boosting, RF based on tree-ensembles, support vector regressor (SVR) as a kernel-based algorithm, and the multilayer perceptron (MLP) as a neural-network approach. We compare the *model fit* of the black-box models to the test data against mean and median as baseline estimators and the intrinsically interpretable white-box approaches MLR and EBM (Table 4).

The figures suggest the following three conclusions. First, the baseline estimators seem to have little to no explanatory power (i.e., explanation of variance in the target variable), which results in comparatively poor model fit. The use of more sophisticated algorithmic approaches, therefore, seems reasonable. Second, the boosting-based learners that require a post hoc explainer display the best results among all models; while the CatBoost shows the lowest error in MAPE, the XGBoost model yields the best results for R^2, and the lowest errors for RMSE and MAE. As the residuals are squared, the RMSE penalizes larger deviations from the actual value stronger than MAE, suggesting that the SVR has (within the category of black-box models) comparatively higher isolated errors in prediction on the test data. For the present case and final model selection for the post hoc explainers, we primarily focus on R^2 for the baseline assessment and additionally on RMSE, as the electricity consumption of a population contains nonnegligible outliers. Hence, we rely on

TABLE 4 | Model fit of the (underlying) models for explanation for the test set.

Explainer approach	Algorithm	R^2	RMSE	MAE	MAPE
Post-hoc explainer (i.e., black-box models)	XGBoost	**0.670**	**9.296**	**6.756**	0.245
	CatBoost	0.642	9.673	6.882	**0.235**
	LightGBM	0.654	9.508	6.961	0.247
	RF	0.587	10.391	7.484	0.256
	SVR	0.502	11.419	7.902	0.267
	MLP	0.584	10.437	7.510	0.242
White-box	EBM	0.613	10.067	7.371	0.269
	MLR	0.535	11.041	8.237	0.300
Baseline	Mean	−0.000	16.174	12.018	0.519
	Median	−0.023	16.356	11.778	0.474

Note: Bold values indicate the best performance for each metric.

XGBoost[2] as the underlying model for the post hoc explainers LIME and SHAP. Third, the results of the white-box models (i.e., EBM and MLR) show, compared with black-box models, a lower model fit. A nonparametric two-sided paired Wilcoxon signed rank test on the residuals of the test data and the best model of both categories confirms these results by displaying a significant difference between the XGBoost and the EBM ($p = 0.036$, $r = 0.113$) but with only a small effect size.

For the following explainer method validation, we use the white-box methods EBM and MLR for forecasting electricity demand (i.e., the "EUI"), as these are method-specific and already intrinsically interpretable methods (i.e., the forecasting model already yields explanations). For the post hoc explainer methods SHAP and LIME, the XGBoost forecasting model serves to predict the EUI. Consequently, all forecasts made in the follow-up analyses for the category of post hoc explainer methods are based on the underlying XGBoost model.

4.4.2 | Consistency

To evaluate internal consistency per method, we divide the data set into $k = 5$ stratified folds in the first step leveraging––compared with model fit––the entire dataset, as this metric focuses on challenging the goodness of an explanation. The folds are similarly distributed regarding the target variable "EUI." We perform Kruskal–Wallis tests to ensure the similarity of the folds and that the stratification works as intended. Our analysis indicates that there is no significant difference in the distribution between folds ($p > 0.940$ for the "one versus all" test per fold). In the second step, we apply each of the explainers to all folds and use an analysis of variances (i.e., a Fligner-Killeen test) to assess how consistent the explanations per feature are between folds. We report a method's internal consistency as the percentage of features for which we find no significant differences in the explanations (significance level $a = 0.1$) between the folds (Figure 2).

Our results show that the ability to recognize nonlinear relationships allows consistent explanations. Thereby, SHAP shows the highest explanation consistency with 100% (i.e., no significant difference for all features across the folds), followed by EBM with 92.9% (i.e., 13 of the 14 features) and LIME with 85.7% (i.e., 12 of the 14 features). Only the linear explanatory method MLR appears with 14.3% to assign attributions with little consistency between the distributions, although the distribution of the data shows no difference, which may be due to the method's outlier sensitivity.

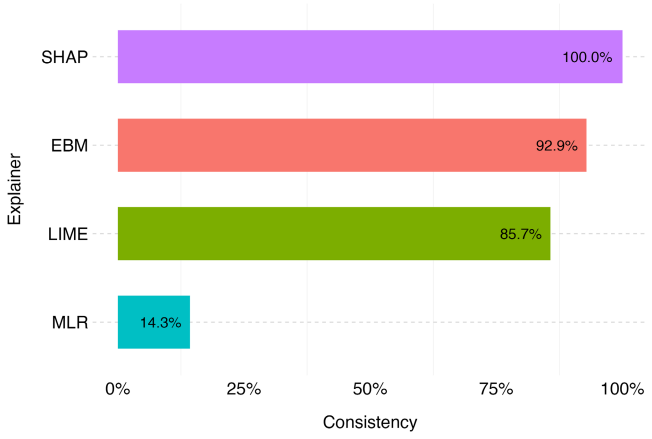


FIGURE 2 | Results of the consistency analysis for the methods SHAP, EBM, LIME, and MLR.

4.4.3 | Stability

Our measurement also considers the entire dataset to obtain *stability* for the explainer methods at hand. Before using Gower's distance to compute dissimilarities between observations, we log-transformed *floorArea*, *numberAdults*, *numberKids*, and *ageBuilding* due to their high positive skewness, and *numberPeaks* due to high negative skewness in the data. Building on the distance matrix, our implementation runs the PAM algorithm multiple times to divide the dataset into similar subsets and to evaluate *stability* for multiple scenarios and each explainer method. For this purpose, we consider a range of two to 50 for k (i.e., medoids), as we observe a stagnant trend for the explainer methods in *stability* for $k > 50$. We fit a local polynomial regression function (LOESS) and the 95% confidence interval of the estimates (shown in gray) over the results to highlight the trend with increasing cluster size (Figure 3).

Our results show that as k increases, *stability* decreases for all methods as the variability of selected features considered as most important increases with a higher number of clusters (i.e., subsets). Overall, we observe the highest stability across subsets for EBM ($M = 0.716$, $SD = 0.066$) followed by SHAP ($M = 0.697$, $SD = 0.065$), MLR ($M = 0.660$, $SD = 0.054$). LIME ($M = 0.634$, $SD = 0.043$), in contrast, results in the lowest stability among the explainer methods, which is consistent with findings from previous research in the context of classification problems (Schlegel et al. 2019; Velmurugan et al. 2021b) and thus strengthens the validity of our approach. The differences for the stability values and cluster sizes ranging from two to 50 are statistically significant for the comparative cases EBM versus SHAP ($p < 0.001$, $r = 0.460$), SHAP versus MLR ($p < 0.001$, $r = 0.870$), and MLR versus LIME ($p = 0.050$, $r = 0.274$). Interestingly, EBM seems to provide more *stable* explanations than all other explainer methods.

4.4.4 | Faithfulness

To assess the *faithfulness* of an explainer method, we blur the values of a feature considered most important with the respective feature mean. The most important feature is determined through the highest assigned feature attribution for the prediction. The reason for the suitability of blurring with the mean is that in this case it can meaningfully assign less influence to the corresponding feature for the model output while, for example, replacing it with zero would here have a specific meaning (e.g., zero occupants in a house or zero floor area). Although we are aware that for the white-box models, the model prediction shifts in the respective other direction (due to the calculation of feature attributions and the inherent transparency regarding the explanation function), this analysis allows us to assess if this also holds for the post hoc

explainer methods that use XGBoost as the underlying forecasting model (hence, the post hoc explainer methods LIME and SHAP are here described as "SHAP + XGBoost" and "LIME + XGBoost").

After replacement, the model outputs change in a desirable direction for all explainer methods when perturbing the most positive and negative impact features, respectively. For the case of the perturbation of features with the highest positive impact (Figure 4a), the model output changes most for EBM ($M = -7.914$, $SD = 5.247$) and SHAP+XGBoost ($M = -7.273$, $SD = 6.935$), and also but less for LIME + XGBoost ($M = -6.884$, $SD = 7.274$) and MLR ($M = -5.320$, $SD = 2.812$). Interestingly, the change in prediction seems to spread less around the mean for MLR compared with the other explainer methods. Likewise, we find a shift in the prediction toward the opposite direction for all explainer methods when the feature values with the most negative impact are replaced by the feature mean (Figure 4b)––however, the change in prediction deviates less. On average, the prediction of the EBM model changes by $M = 5.060$ ($SD = 4.126$), for LIME + XGBoost by $M = 5.535$ ($SD = 5.591$), for MLR by $M = 6.200$ ($SD = 2.278$), and for SHAP + XGBoost by $M = 6.372$ ($SD = 4.965$). We conduct two-sided paired Wilcoxon signed rank tests for both perturbation directions and the most contradictory explainer methods (i.e., EBM vs. MLR for the most positive impact feature perturbed and EBM vs. SHAP + XGBoost for the opposite direction). We observe for both methods a significant difference for the change in prediction with a large effect size when perturbing the most positive ($p < 0.001$, $r = 0.721$) and a small effect size when perturbing the most negative impact feature ($p < 0.001$, $r = 0.250$). We conclude that all methods can be said to be *faithful* in our case, although there is a large difference between explainer methods when the most positive feature is perturbed.

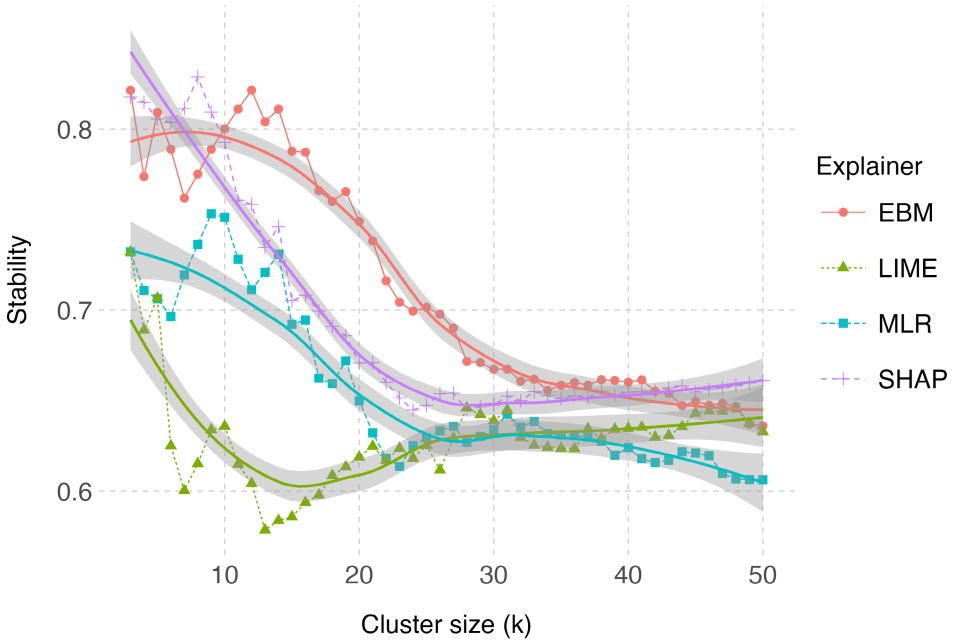
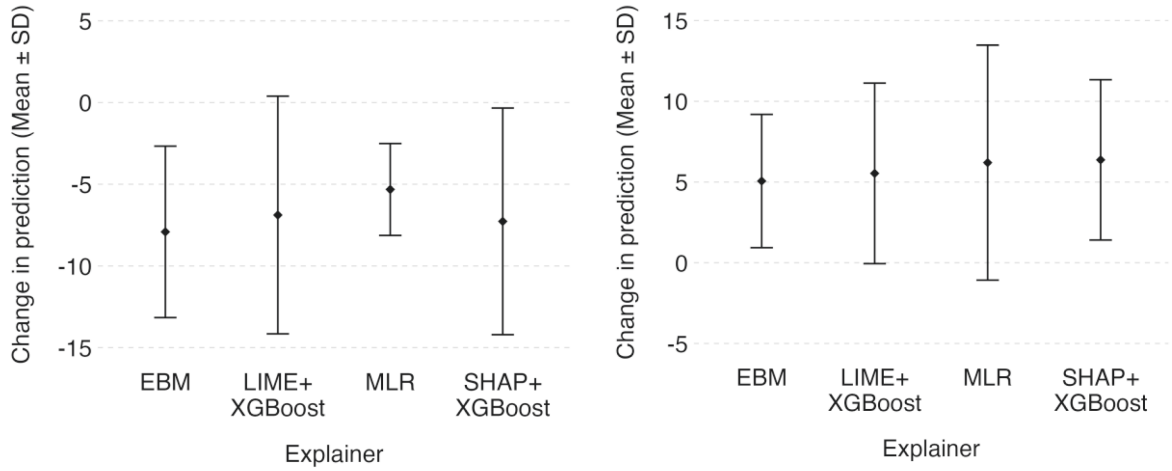


FIGURE 3 | Stability results of explainer methods over varying cluster size (with LOESS trend estimates and its 95% confidence interval in gray).



(a) Most positive impact features perturbed *(b) Most negative impact feature perturbed*

FIGURE 4 | Change in prediction after feature perturbation according to the highest impact feature.

4.5 | Summary of the Results of the Demonstration Case

We demonstrated the applicability of the FEXVAL approach by validating and comparing the explainer methods EBM, LIME, MLR, and SHAP by employing the case of energy benchmarking using a real-world dataset from 1262 residential customers. To provide a holistic overview of the results, we applied min-max normalization across all metrics and explainer methods (for the metric *faithfulness*, we calculated the average of the positive and absolute negative impacts). In doing so, we first validated the methods by assessing the results for the categories "post-hoc explainer methods" and "white-box explainer methods" separately (Figure 5).

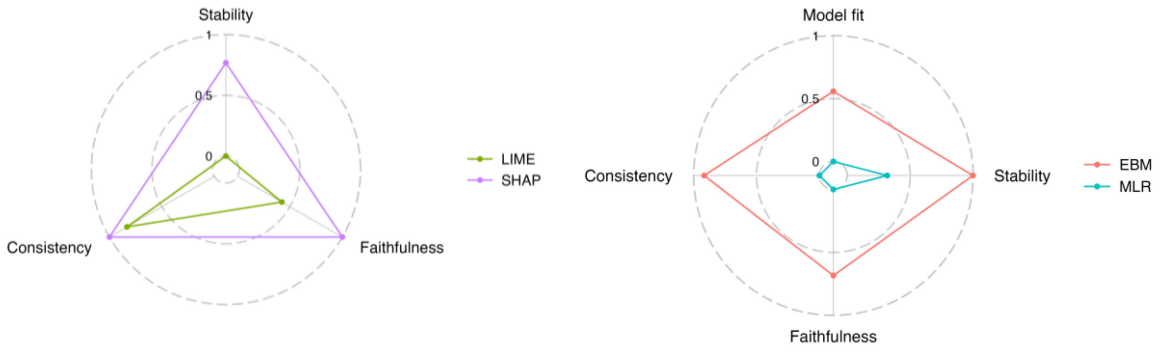
Given that the post-hoc explainer methods are based on the same underlying ML model, we omit *model fit* for their assessment (Figure 5a). Overall, we find the following: When choosing post-hoc explainers, we find that SHAP is superior to LIME across all metrics. Interestingly, SHAP seems to offer the highest faithfulness and consistency across all methods and yields comparatively high stability; LIME, in contrast, displays the worst performance in terms of stability. A similar picture regarding the dominance of one approach arises for the case at hand when one decides to employ white-box explainer methods (Figure 5b): The EBM approach is superior to the linear statistical approach MLR across all metrics; MLR performs worse regarding model fit, consistency, and faithfulness. In addition, EBM appears to have the highest stability across all explainer methods.

Given that FEXVAL is method-agnostic, it allows for a trade-off calculation across post-hoc explainer methods and white-box approaches, with the metrics being aggregated using the geometric mean. Assuming equal weights and setting $\eta = 0.01$ for the overall performance metric P, we find that SHAP yields the best results in our case ($P = 0.930$), followed by EBM ($P = 0.749$) and LIME ($P = 0.418$). The linear explainer method MLR provides the lowest overall performance ($P = 0.220$). Notably, EBM, as a white-box approach, ranks first for stability and second to third for all other metrics. This result allows for the following conclusion for the case and dataset at hand: Among post-hoc explainer methods, SHAP appears to be particularly effective for explaining model forecasts in our case; this results is in line with previous studies on classification tasks, which found SHAP to be more powerful than LIME in terms of stability (Schlegel et al. 2019). Interestingly, our study shows that, in the present case, the EBM outperforms traditional and linear statistical models across the metrics *stability* and *faithfulness*. Hence, when developers integrate a white-box approach into a forecasting system for the case at hand, the EBM serves as an adequate intermediate solution, performing comparably to SHAP while also offering inherent interpretability.

5 | Discussion

Data-driven explanations embedded in forecasting systems promise significant benefits for decision-making processes (Harl et al. 2020; Shin 2021). Given the importance of explanations related to forecasts, various categories of methods have been employed. Starting from inherently interpretable and mostly linear models (G. E. Phillips-Wren and Forgionne 2002), the last decade has seen increasing use of XAI methods that make the nonlinear relationships that complex ML discovers in the data comprehensible to decision-makers (Moreira et al. 2021). The type of explainer method that is integrated into forecasting systems, but also the suitability and fit of these for a specific problem, exerts a direct influence on the sociotechnical system (Miller 2019). Hence, as the choice of explainer method can have a large influence on the resulting decisions (Herm 2023), the functional validation of these methods in the early stages of system development can benefit developers, stakeholders, and, ultimately, decision-makers. While there are isolated proposals for metrics in classification problems (Velmurugan et al. 2021a, 2021b), there is a notable lack of approaches for numerical prediction tasks. In our paper, we address this issue by introducing FEXVAL, which allows for a functionally grounded validation of explainer methods within but also between method categories.

Our demonstration shows that FEXVAL is an effective tool for validating and comparing different explainer methods. For the case of energy benchmarking, we find a confirmation of the general assumption that black-box-based models are superior to linear statistical models in terms of *model fit* due to their ability to reflect nonlinearities (Breiman 2001; Shmueli 2010). This result reinforces the existence of the accuracy-interpretability trade-off (Barredo Arrieta et al. 2020) when it comes to a comparison with rather simple linear models; however, this does not hold when the EBM comes into play, which offers a comparable *model fit* to black-box approaches with only a small difference in the effect



(a) Post-hoc explainer methods results *(b) White-box explainer methods results*

FIGURE 5 | Summary of results for all metrics per explainer method category.

size. Beyond this trade-off, we observe that the simplicity of MLR, which leads to an inadequately fitted model (James et al. 2013), comes with additional drawbacks: The MLR model also provides less stable and consistent explanations. In summary, SHAP seems to be the currently superior approach in the field of post hoc explainer methods considering the case at hand. For white-box approaches, EBM seems dominant in the present case study.

In summary, our paper makes three contributions: First, independent of a specific application domain, the FEXVAL approach *offers a means to compare and validate explainer methods for numerical forecasts on a functional-grounded level*. Researchers and practitioners can apply our approach for baseline validation and comparison of the quality of quantitative explainer methods on multiple levels. To the best of our knowledge, this is the first comprehensive validation approach that defines such metrics for numerical forecasting. Explainer method validation using FEXVAL has two main advantages over testing the explanations at the application or human-grounded level: A functionally grounded validation is closer in time to the actual technical implementation and can, therefore, already take place during the development stage of a forecasting system. Predefined threshold values might be implemented in automatic software tests, realizing continuous quality assurance. A further downstream advantage is that the functional-grounded validations are performed without human involvement and are therefore of no subjective nature and cost-intensive, as is often the case with complex empirical laboratory and field studies. Second, we *defined computational rules for all metrics* included in FEXVAL for their application in numerical forecasting. To date, such computational rules have only been known for classification. Third, we *demonstrated the approach using energy benchmarking as a use case*. We chose energy benchmarking because it is a well-established and common use case for forecasting systems, where both linear statistical models and XAI methods have been effectively applied to explain energy consumption patterns that are related to user characteristics and behavior (Arjunan et al. 2020).

While our findings demonstrate the effectiveness of FEXVAL in the context of energy benchmarking, it is important to note that the generalizability of these results only extends to some degree beyond this specific use case. Previous research has observed similar outcomes when employing SHAP and LIME in classification tasks (Schlegel et al. 2019; Velmurugan et al. 2021a), suggesting that these explainer methods may exhibit consistent behavior across different types of tasks and domains. In addition, the relationships in the data are given by natural phenomena, and various explainer methods are already in practical use to support the decision-making of stakeholders (e.g., Huebner et al. 2015; Arjunan et al. 2020). We therefore assume that our results are generalizable to a certain extent. However, we emphasize that each dataset and scenario presents unique challenges, necessitating a case-specific analysis using the FEXVAL approach. As such, while FEXVAL provides a robust approach for explainer method validation, it is essential for researchers and practitioners to conduct thorough validation tailored to the datasets and forecasting objectives at hand to ensure the effectiveness of the explainer method in the specific context.

Despite our best efforts, our work has several limitations that open avenues for future research. For example, the FEXVAL approach focuses solely on feature attributions, while other types of explanation, such as counterfactual explanations, are also frequently applied in the field (Fernández-Loría et al. 2022). Furthermore, our approach is currently limited to the contributions of individual features to compare the respective explainer methods meaningfully. Other aspects, such as feature interactions, are not yet included. Future research on functionally grounded validations may, therefore, also cover other explanation types, (automated) feature interactions, or aspects of inferential statistics like empirical tests for causality. Additionally, FEXVAL consists at present of carefully and deliberately selected metrics that belong to the category of functionally grounded validation metrics (see Table 1). However, defining metrics for functionally grounded validation and comparison of explainer methods is an infinite space, as multiple further scenarios for measuring method quality are conceivable beyond those defined by current literature. Hence, our approach also allows for extensions by integrating novel metrics and associated calculation rules. We encourage other scholars in forecasting research to extend FEXVAL where appropriate and test the approach for further cases and possible newly proposed metrics.

6 | Conclusion

Explanations related to numeric forecasts constitute important insights for managerial decision-making (Shin 2021). Given the importance of explanations, scholars have begun to assess the quality of explanations, yet research dominantly focuses on subjective application and user-oriented evaluations, classification tasks (Bauer et al. 2023; Herm 2023), and in technical terms either on XAI (Alvarez Melis and Jaakkola 2018; Velmurugan et al. 2021b) or inherently interpretable linear statistical method validations (Mészáros and Rapcsák 1996). Our work addresses this research gap by developing FEXVAL, an approach that allows for the validation and comparison of explainer methods of different categories (i.e., traditional (linear) statistical methods and more recent post hoc and white-box explainer methods). The outlined approach of this paper draws on metrics that have already been conceptually defined in the literature; however, for such, no concrete computational rules were known, in particular not for numerical forecasting. The demonstration shows that FEXVAL is well suited for the purpose of a method-independent functionally grounded validation and comparison of explainer methods using the deliberately chosen metrics. Our approach provides forecasting research and practice with the opportunity for a technical baseline validation and comparison of explainer methods for numerical forecasting tasks and opens avenues for future research. This enables (a) firms to validate explainer methods in the development process of decision systems in a cost-efficient manner, and (b) forecasting systems research to validate and compare explainer methods on a functionally grounded level.

Acknowledgments

Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in the Irish Social Science Data Archive at https://www.ucd.ie/issda/data/commissionforenergyregulationcer, reference number 0012-00. Upon request, we can also provide the code of our analyses.

Endnotes

[1] Literature has also referred to the metrics *model fit* as *predictive accuracy* and *faithfulness* as *fidelity* (Robnik-Šikonja and Bohanec 2018; Schwalbe and Finzel 2023).

[2] Final XGBoost parameters: *n_estimators* = *250*, *max_depth* = *3*, *learning_rate* = *0.05*, *min_child_weight* = *5*, *gamma* = *0*, *subsample* = *0.4*, *reg_alpha* = *50*, *reg_lambda* = *0.1*.

References

Abdul, A., J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. 2018. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18. ACM (Association for Computing Machinery). https://doi.org/10.1145/3173574.3174156.

Adadi, A., and M. Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Alvarez Melis, D., and T. Jaakkola. 2018. "Towards Robust Interpretability With Self-Explaining Neural Networks." *Advances in Neural Information Processing Systems* 31: 7786–7795. https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

Antunes, P., V. Herskovic, S. F. Ochoa, and J. A. Pino. 2008. "Structuring Dimensions for Collaborative Systems Evaluation." *ACM Computing Surveys* 44, no. 2: 1–28. https://doi.org/10.1145/2089125.2089128.

Arjunan, P., K. Poolla, and C. Miller. 2020. "EnergyStar++: Towards More Accurate and Explanatory Building Energy Benchmarking." *Applied Energy* 276: 115413. https://doi.org/10.1016/j.apenergy.2020.115413.

Armstrong, J. S., and F. Collopy. 1992. "Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons." *International Journal of Forecasting* 8, no. 1: 69–80.

Aspers, P. 2018. "Forms of Uncertainty Reduction: Decision, Valuation, and Contest." *Theory and Society* 47, no. 2: 133–149. https://doi.org/10.1007/s11186-018-9311-0.

Bansal, G., T. Wu, J. Zhou, et al. 2021. "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.ACM (Association for Computing Machinery). https://doi.org/10.1145/3411764.3445717.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bauer, K., O. Hinz, W. van der Aalst, and C. Weinhardt. 2021. "Expl(AI)n It to Me—Explainable AI and Information Systems Research." *Business & Information Systems Engineering* 63, no. 2: 79–82. https://doi.org/10.1007/s12599-021-00683-2.

Bauer, K., M. Von Zahn, and O. Hinz. 2023. "Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing." *Information Systems Research* 34: 1582–1602. https://doi.org/10.1287/isre.2023.1199.

Borenstein, D. 1998. "Towards a Practical Method to Validate Decision Support Systems." *Decision Support Systems* 23, no. 3: 227–239. https://doi.org/10.1016/S0167-9236(98)00046-3.

Branley-Bell, D., R. Whitworth, and L. Coventry. 2020. "User Trust and Understanding of Explainable AI: Exploring Algorithm Visualisations and User Biases." In *Human-Computer Interaction. Human Values and Quality of Life*, edited by M. Kurosu, vol. 12183, 382–399. Springer International Publishing. https://doi.org/10.1007/978-3-030-49065-2_27.

Breiman, L. 2001. "Statistical Modeling: The two Cultures (With Comments and a Rejoinder by the Author)." *Statistical Science* 16, no. 3: 199–231.

Carvalho, D. V., E. M. Pereira, and J. S. Cardoso. 2019. "Machine Learning Interpretability: A Survey on Methods and Metrics." *Electronics* 8, no. 8: 832.

Central Statistics Office Ireland. 2010. "Planning Permissions Granted for new Houses and Apartments (BHQ05)." https://data.cso.ie/table/BHQ05.

Chen, X., Y. Wang, H. Zhang, and J. Wang. 2024. "A Novel Hybrid Forecasting Model With Feature Selection and Deep Learning for Wind Speed Research." *Journal of Forecasting* 43: 1682–1705. https://doi.org/10.1002/for.3098.

Chung, W. 2011. "Review of Building Energy-Use Performance Benchmarking Methodologies." *Applied Energy* 88, no. 5: 1470–1479. https://doi.org/10.1016/j.apenergy.2010.11.022.

Commission for Energy Regulation. 2011. *Electricity Smart Metering Technology Trials Findings Report (CER11080b)*. Irish Social Science Data Archive.

Conover, W. J., M. E. Johnson, and M. M. Johnson. 1981. "A Comparative Study of Tests for Homogeneity of Variances, With Applications to the Outer Continental Shelf Bidding Data." *Technometrics* 23, no. 4: 351–361. https://doi.org/10.1080/00401706.1981.10487680.

Corrente, S., and M. Tasiou. 2023. "A Robust TOPSIS Method for Decision Making Problems With Hierarchical and Non-Monotonic Criteria." *Expert Systems with Applications* 214: 119045. https://doi.org/10.1016/j.eswa.2022.119045.

De Castro Moraes, T., X. Yuan, and E. P. Chew. 2024. "Hybrid Convolutional Long Short-term Memory Models for Sales Forecasting in Retail." *Journal of Forecasting* 43: for.3073. https://doi.org/10.1002/for.3073.

Dol, K., and M. Haffner. 2010. *Housing Statistics in the European Union 2010*. Delft University of Technology.

Doshi-Velez, F., & B. Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [cs, Stat]. http://arxiv.org/abs/1702.08608.

Doszyń, M. 2019. "Intermittent Demand Forecasting in the Enterprise: Empirical Verification." *Journal of Forecasting* 38, no. 5: 459–469. https://doi.org/10.1002/for.2575.

Du, M., N. Liu, F. Yang, S. Ji, and X. Hu. 2019. "On Attribution of Recurrent Neural Network Predictions via Additive Decomposition." In *The World Wide web Conference on – WWW'19*, 383–393. ACM (Association for Computing Machinery). https://doi.org/10.1145/3308558.3313545.

Ducharme, C., B. Agard, and M. Trépanier. 2024. "Improving Demand Forecasting for Customers With Missing Downstream Data in Intermittent Demand Supply Chains With Supervised Multivariate Clustering." *Journal of Forecasting* 43: 1661–1681.

European Commission. 2012. *Energy Roadmap 2050*. Publications Office of the European Union. https://data.europa.eu/doi/10.2833/10759.

Fan, G., Y. Guo, J. Zheng, and W. Hong. 2020. "A Generalized Regression Model Based on Hybrid Empirical Mode Decomposition and Support Vector Regression With Back-Propagation Neural Network for Mid-Short-Term Load Forecasting." *Journal of Forecasting* 39, no. 5: 737–756. https://doi.org/10.1002/for.2655.

Fernández-Loría, C., F. Provost, and X. Han. 2022. "Explaining Data-Driven Decisions Made by AI Systems: The Counterfactual Approach." *MIS Quarterly* 46, no. 3: 1635–1660. https://doi.org/10.25300/MISQ/2022/16749.

Finlay, P. N. 1994. *Introducing Decision Support Systems (1. Publ.)*. NCC Blackwell.

Fischer, H., Á. Blanco-FERNÁndez, and P. Winker. 2016. "Predicting Stock Return Volatility: Can We Benefit From Regression Models for Return Intervals?" *Journal of Forecasting* 35, no. 2: 113–146. https://doi.org/10.1002/for.2371.

Fleming, P. J., and J. J. Wallace. 1986. "How Not to Lie With Statistics: The Correct Way to Summarize Benchmark Results." *Communications of the ACM* 29, no. 3: 218–221. https://doi.org/10.1145/5666.5673.

Fligner, M. A., and T. J. Killeen. 1976. "Distribution-Free Two-Sample Tests for Scale." *Journal of the American Statistical Association* 71, no. 353: 210–213.

Goodwin, P., and G. Wright. 2014. *Decision Analysis for Management Judgment*. 5th ed. Wiley.

Gower, J. C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics* 27: 857–871.

Gregor, S., and I. Benbasat. 1999. "Explanations From Intelligent Systems: Theoretical Foundations and Implications for Practice." *MIS Quarterly* 23, no. 4: 497. https://doi.org/10.2307/249487.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2018. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51, no. 5: 93:1. https://doi.org/10.1145/3236009.

Harl, M., S. Weinzierl, M. Stierle, and M. Matzner. 2020. "Explainable Predictive Business Process Monitoring Using Gated Graph Neural Networks." *Journal of Decision Systems* 29, no. sup1: 312–327. https://doi.org/10.1080/12460125.2020.1780780.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7.

Herm, L.-V. 2023. "Impact of Explainable AI on Cognitive Load: Insights From an Empirical Study." In *Proceedings of the 31st European Conference on Information Systems (ECIS 2023)*. Association for Information Systems (AIS).

Hudon, A., T. Demazure, A. Karran, P.-M. Léger, and S. Sénécal. 2021. "Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence." In *Proceedings NeuroIS Retreat 2021*. Springer.

Huebner, G. M., I. Hamilton, Z. Chalabi, D. Shipworth, and T. Oreszczyn. 2015. "Explaining Domestic Energy Consumption—The Comparative Contribution of Building Factors, Socio-Demographics, Behaviours and Attitudes." *Applied Energy* 159: 589–600. https://doi.org/10.1016/j.apenergy.2015.09.028.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning (Bd. 103)*. Springer. http://link.springer.com/10.1007/978-1-4614-7138-7.

Jiang, H., and W. Zheng. 2022. "Deep Learning With Regularized Robust Long- and Short-Term Memory Network for Probabilistic Short-Term Load Forecasting." *Journal of Forecasting* 41, no. 6: 1201–1216. https://doi.org/10.1002/for.2855.

Kaufman, L., and P. J. Rousseeuw. 1990. "Partitioning Around Medoids (Program PAM)." *Finding Groups in Data: An Introduction to Cluster Analysis* 344: 68–125.

Kim, S., and H. Kim. 2016. "A New Metric of Absolute Percentage Error for Intermittent Demand Forecasts." *International Journal of Forecasting* 32, no. 3: 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003.

Kraus, M., D. Tschernutter, S. Weinzierl, and P. Zschech. 2023. "Interpretable Generalized Additive Neural Networks." *European Journal of Operational Research* 317, no. 2: 303–316. https://doi.org/10.1016/j.ejor.2023.06.032.

Lai, V., H. Liu, and C. Tan. 2020. ""Why Is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. ACM (Association for Computing Machinery). https://doi.org/10.1145/3313831.3376873.

Lakkaraju, H., S. H. Bach, and J. Leskovec. 2016. "Interpretable Decision Sets: A Joint Framework for Description and Prediction." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. Association for Computing Machinery (ACM). https://doi.org/10.1145/2939672.2939874.

Lee, S., S.-H. Kim, Y.-H. Hung, H. Lam, Y.-A. Kang, and J. S. Yi. 2016. "How Do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking." *IEEE Transactions on Visualization and Computer Graphics* 22, no. 1: 499–508. https://doi.org/10.1109/TVCG.2015.2467195.

Lim, B., S. Ö. Arık, N. Loeff, and T. Pfister. 2021. "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting." *International Journal of Forecasting* 37, no. 4: 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012.

Liu, J., C. Li, P. Ouyang, J. Liu, and C. Wu. 2023. "Interpreting the Prediction Results of the Tree-Based Gradient Boosting Models for Financial Distress Prediction With an Explainable Machine Learning Approach." *Journal of Forecasting* 42, no. 5: 1112–1137. https://doi.org/10.1002/for.2931.

Liu, Y., F. Huang, L. Ma, Q. Zeng, and J. Shi. 2024. "Credit Scoring Prediction Leveraging Interpretable Ensemble Learning." *Journal of Forecasting* 43, no. 2: 286–308. https://doi.org/10.1002/for.3033.

Lou, Y., R. Caruana, and J. Gehrke. 2012. "Intelligible Models for Classification and Regression." In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 150–158. Association for Computing Machinery (ACM). https://doi.org/10.1145/2339530.2339556.

Lou, Y., R. Caruana, J. Gehrke, and G. Hooker. 2013. "Accurate Intelligible Models With Pairwise Interactions." In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 623–631. Association for Computing Machinery (ACM). https://doi.org/10.1145/2487575.2487579.

Lundberg, S. M., G. Erion, H. Chen, et al. 2020. "From Local Explanations to Global Understanding With Explainable AI for Trees." *Nature Machine Intelligence* 2, no. 1: 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Lundberg, S. M., & S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions."

Meske, C., E. Bunde, J. Schneider, and M. Gersch. 2020. "Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities." *Information Systems Management* 39, no. 1: 53–63. https://doi.org/10.1080/10580530.2020.1849465.

Mészáros, C., and T. Rapcsák. 1996. "On Sensitivity Analysis for a Class of Decision Systems." *Decision Support Systems* 16, no. 3: 231–240. https://doi.org/10.1016/0167-9236(95)00012-7.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights From the Social Sciences." *Artificial Intelligence* 267: 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Mohana Chelvan, P., and K. Perumal. 2016. "A Survey on Feature Selection Stability Measures." *International Journal of Computer and Information Technology* 5, no. 1: 98–103.

Molnar, C. 2019. "Interpretable Machine Learning—A Guide for Making Black box Models Explainable." https://christophm.github.io/interpretable-ml-book/.

Moreira, C., Y.-L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, and P. Bruza. 2021. "LINDA-BN: An Interpretable Probabilistic Approach for Demystifying Black-Box Predictive Models." *Decision Support Systems* 150: 113561. https://doi.org/10.1016/j.dss.2021.113561.

Nogueira, S., K. Sechidis, and G. Brown. 2018. "On the Stability of Feature Selection Algorithms." *Journal of Machine Learning Research* 18, no. 174: 1–54.

Nori, H., S. Jenkins, P. Koch, and R. Caruana. 2019. "InterpretML: A Unified Framework for Machine Learning Interpretability (arXiv:1909.09223)." arXiv. https://doi.org/10.48550/arXiv.1909.09223.

O'Keefe, R. M., and D. E. O'Leary. 1993. "Expert System Verification and Validation: A Survey and Tutorial." *Artificial Intelligence Review* 7, no. 1: 3–42. https://doi.org/10.1007/BF00849196.

O'Leary, D. E. 1987. "Validation of Expert Systems—With Applications to Auditing and Accounting Expert Systems*." *Decision Sciences* 18, no. 3: 468–486. https://doi.org/10.1111/j.1540-5915.1987.tb01536.x.

O'Leary, T. J., M. Goul, K. E. Moffitt, and A. E. Radwan. 1990. "Validating Expert Systems." *IEEE Expert* 5, no. 3: 51–58. https://doi.org/10.1109/64.54673.

Önkal, D., P. Goodwin, M. Thomson, S. Gönül, and A. Pollock. 2009. "The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Adjustments." *Journal of Behavioral Decision Making* 22, no. 4: 390–409. https://doi.org/10.1002/bdm.637.

Palmer, K. L., and M. Walls. 2015. "Does Information Provision Shrink the Energy Efficiency Gap? A Cross-City Comparison of Commercial Building Benchmarking and Disclosure Laws." *SSRN*. https://doi.org/10.2139/ssrn.2622692.

Peters, G. 2001. "A Linear Forecasting Model and Its Application to Economic Data." *Journal of Forecasting* 20, no. 5: 315–328. https://doi.org/10.1002/for.795.

Phillips-Wren, G. E., and G. A. Forgionne. 2002. "Advanced Decision Making Support Using Intelligent Agent Technology." *Journal of Decision Systems* 11, no. 2: 165–184. https://doi.org/10.3166/jds.11.165-184.

Phillips-Wren, G., M. Mora, G. A. Forgionne, and J. N. D. Gupta. 2009. "An Integrative Evaluation Framework for Intelligent Decision Support Systems." *European Journal of Operational Research* 195, no. 3: 642–652. https://doi.org/10.1016/j.ejor.2007.11.001.

Rai, A. 2020. "Explainable AI: From Black Box to Glass Box." *Journal of the Academy of Marketing Science* 48, no. 1: 137–141. https://doi.org/10.1007/s11747-019-00710-5.

Rao, R. B., G. Fung, and R. Rosales. 2008. "On the Dangers of Cross-Validation. An Experimental Evaluation." In *Proceedings of the 2008 SIAM International Conference on Data Mining*, 588–596. Society for Industrial and Applied Mathematics (SIAM).

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

Roach, C., R. Hyndman, and S. Ben Taieb. 2021. "Non-Linear Mixed-Effects Models for Time Series Forecasting of Smart Meter Demand." *Journal of Forecasting* 40, no. 6: 1118–1130. https://doi.org/10.1002/for.2750.

Robinson, C., B. Dilkina, J. Hubbs, et al. 2017. "Machine Learning Approaches for Estimating Commercial Building Energy Consumption." *Applied Energy* 208: 889–904. https://doi.org/10.1016/j.apenergy.2017.09.060.

Robnik-Šikonja, M., and M. Bohanec. 2018. "Perturbation-Based Explanations of Prediction Models." In *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, 159–175. Springer International Publishing. https://doi.org/10.1007/978-3-319-90403-0_9.

Saaty, T. L. 2008. "Decision Making With the Analytic Hierarchy Process." *International Journal of Services Sciences* 1, no. 1: 83–98.

Santin, O. G., L. Itard, and H. Visscher. 2009. "The Effect of Occupancy and Building Characteristics on Energy Use for Space and Water Heating in Dutch Residential Stock." *Energy and Buildings* 41, no. 11: 1223–1232. https://doi.org/10.1016/j.enbuild.2009.07.002.

Schlegel, U., H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim. 2019. "Towards a Rigorous Evaluation of XAI Methods on Time Series." In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 4197–4201. IEEE. https://doi.org/10.1109/ICCVW.2019.00516.

Schwalbe, G., and B. Finzel. 2023. "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts." *Data Mining and Knowledge Discovery* 38, no. 5: 3043–3101. https://doi.org/10.1007/s10618-022-00867-8.

Shin, D. 2021. "The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI." *International Journal of Human-Computer Studies* 146: 102551. https://doi.org/10.1016/j.ijhcs.2020.102551.

Shmueli, G. 2010. "To Explain or to Predict?" *Statistical Science* 25, no. 3: 289–310. https://doi.org/10.1214/10-STS330.

Shmueli, G., and O. R. Koppius. 2011. "Predictive Analytics in Information Systems Research." *MIS Quarterly* 35, no. 3: 553–572.

Štrumbelj, E., and I. Kononenko. 2014. "Explaining Prediction Models and Individual Predictions With Feature Contributions." *Knowledge and Information Systems* 41, no. 3: 647–665. https://doi.org/10.1007/s10115-013-0679-x.

Štrumbelj, E., I. Kononenko, and M. R. Šikonja. 2009. "Explaining Instance Classifications With Interactions of Subsets of Feature Values." *Data & Knowledge Engineering* 68, no. 10: 886–904.

Swartout, W. R. 1983. "XPLAIN: A System for Creating and Explaining Expert Consulting Programs." *Artificial Intelligence* 21, no. 3: 285–325. https://doi.org/10.1016/S0004-3702(83)80014-9.

Theising, E., D. Wied, and D. Ziggel. 2023. "Reference Class Selection in Similarity-Based Forecasting of Corporate Sales Growth." *Journal of Forecasting* 42, no. 5: 1069–1085. https://doi.org/10.1002/for.2927.

Velmurugan, M., C. Ouyang, C. Moreira, and R. Sindhgatta. 2021a. "Evaluating Fidelity of Explainable Methods for Predictive Process Analytics." In *Intelligent Information Systems*, edited by S. Nurcan and A. Korthaus, vol. 424, 64–72. Springer International Publishing. https://doi.org/10.1007/978-3-030-79108-7_8.

Velmurugan, M., C. Ouyang, C. Moreira, and R. Sindhgatta. 2021b. "Evaluating Stability of Post-Hoc Explanations for Business Process Predictions." In *Service-Oriented Computing*, edited by H. Hacid, O. Kao, M. Mecella, N. Moha, and H. Paik, vol. 13121, 49–64. Springer International Publishing. https://doi.org/10.1007/978-3-030-91431-8_4.

Wang, J. C. 2019. "Analysis of Energy Use Intensity and Greenhouse Gas Emissions for Universities in Taiwan." *Journal of Cleaner Production* 241: 118363. https://doi.org/10.1016/j.jclepro.2019.118363.

Wastensteiner, J., T. Weiss, F. Haag, and K. Hopf. 2021. "Explainable AI for Tailored Electricity Consumption Feedback—An Experimental

Evaluation of Visualizations." In *Proceedings of the 29th European Conference on Information Systems (ECIS 2021)*, Marrakech, Morocco (virtual). Association for Information Systems (AIS).

Wu, B., Z. Wang, and L. Wang. 2024. "Interpretable Corn Future Price Forecasting With Multivariate Time Series." *Journal of Forecasting* 43: 3099. https://doi.org/10.1002/for.3099.

Yeomans, M., A. Shah, S. Mullainathan, and J. Kleinberg. 2019. "Making Sense of Recommendations." *Journal of Behavioral Decision Making* 32, no. 4: 403–414. https://doi.org/10.1002/bdm.2118.

Zhang, L., M. Z. Abedin, and Z. Liu. 2024. "Incorporating Media News to Predict Financial Distress: Case Study on Chinese Listed Companies." *Journal of Forecasting* 43: 1374–1398. https://doi.org/10.1002/for.3089.

Zhang, R., Z. Tian, K. J. McCarthy, X. Wang, and K. Zhang. 2023. "Application of Machine Learning Techniques to Predict Entrepreneurial Firm Valuation." *Journal of Forecasting* 42, no. 2: 402–417. https://doi.org/10.1002/for.2912.

Zschech, P., S. Weinzierl, N. Hambauer, S. Zilker, and M. Kraus. 2022. "GAM(E) Changer or Not? An Evaluation of Interpretable Machine Learning Models Based on Additive Model Constraints." In *Proceedings of the 30th European Conference on Information Systems (ECIS 2022)*, Timișoara, Romania. Association for Information Systems (AIS).