This paper is a pre-print of a paper submitted to and accepted for publication in the Conference Proceedings of the **CIRED 2026 Brussels Workshop on Implementing Successful Innovation in Distribution Networks** and is subject to the Organizers' Copyright. The copy of record will be available at The IET Digital Library.

A TEST AND VALIDATION FRAMEWORK FOR GRID DATA ANALYTICS WITH AN APPLICATION TO CONGESTION FORECASTING

Gertjan Kok[1*], Helko van den Brom[1], Ties van der Heijden[2], Bart Pleiter[3]

[1]VSL B.V., Delft, the Netherlands
[2]TU Delft, Delft, the Netherlands
[3]Alliander N.V., Arnhem, the Netherlands
**gkok@vsl.nl*

Keywords: VALIDATION, UNCERTAINTY, DATA ANALYTICS, CONGESTION FORECASTING, METROLOGY

Abstract

In the light of potential regulatory requirements arising from the EU's AI act regarding methos based on artificial intelligence (AI) used in the context of critical infrastructure, as well as to facilitate the communication between suppliers and users of such software, an agreed uniform framework for testing and validating grid data analytics is highly desirable. The European research project 'Metrology for reliable grid data analytics' is developing such a standardized framework. This paper presents a first outline of the framework and how it could be applied for validating congestion forecasting methods. An important observation is that in particular for machine learning (ML) based models having dynamic and high-dimensional input spaces it is insufficient to verify software performance in the classical way only by testing it once for a limited set of inputs and outputs. The draft framework consists of various parts that together aim to lead to trustworthy software including a rigorous uncertainty statement. While it can be a considerable effort to address all parts, such an in-depth validation approach is essential for assuring continued reliability of grid data analytics. This paper focuses on congestion forecasting as a representative high-impact use case.

1 Introduction

An essential part of short-term operational planning in distribution grids consists of forecasting net power loads, in short 'loads', at specific grid nodes up to two days in advance, preventing overloading grid assets. Often a data-driven AI-based approach is employed. An example of a framework facilitating short-term energy forecasts is given by the LF-Energy OpenSTEF project [1], in which the Dutch DSO Alliander is strongly involved.

Such short-term forecasts of at most 48 hours in advance possess a considerable uncertainty, which needs to be properly defined and quantified, and, more generally, the data analysis methods need to be properly tested and validated. This applies not only to forecasting methods, but also to algorithms aiming to detect various types of abnormal grid events.

There are many different approaches for testing data analysis methods. However, in particular in light of potential regulatory requirements arising from the EU's AI act regarding AI methods used in the context of critical infrastructure, as well as to facilitate the communication between suppliers and users of such software, an agreed uniform framework for testing and validating grid data analytics is highly desirable.

The European research project 'Metrology for Reliable Grid Data Analytics' (GridData) [2], that started in June 2025, aims to design such a uniform test framework. The project is investigating several grid data analysis methods for five different grid phenomena (congestion forecasting, early detection of frequency events, sub-synchronous oscillations, power quality disturbances and grid asset faults). The work involves the usage of quality-assured real grid measurement data, supplemented by simulated data for each of the studied phenomena, allowing for a versatile way of testing.

The proposed testing of the methods includes a metrology-based quantification of their uncertainty, determination of the sensitivity of the methods to distribution shifts in the input data, as well as the application of techniques from the explainable AI field. Based on the work for specific use cases, a general test and validation framework will be abstracted.

In this paper some initial ideas and considerations with respect to the key elements of this framework are presented, guided by an example application of congestion forecasting, which is a method based on ML learning.

2 Test and validation framework

The test and validation framework currently being developed by the GridData project aims to assure the reliability of grid data analytics, be as generally applicable as possible, be based on best practices from literature and include a structured metrological approach similar to evaluating measurement uncertainty for which the 'Guide to the Expression of Uncertainty in Measurement' (GUM) [3] is the leading document.

In the use case of congestion forecasting the load at a grid node has to be predicted. In our case we considered load prediction for the next day with a 15-minutes resolution. To this purpose,

an ML model was trained using three months of historical data of load measurements at the node of interest, weather forecasts of about 10 weather variables (e.g., temperature, solar radiation, windspeed), typical user profiles [4] and day ahead electricity prices (EPEX spot NL). In the model under consideration, predictions are performed using two weeks of historical data. An example of a load pattern with congestion limits is shown in Figure 1.

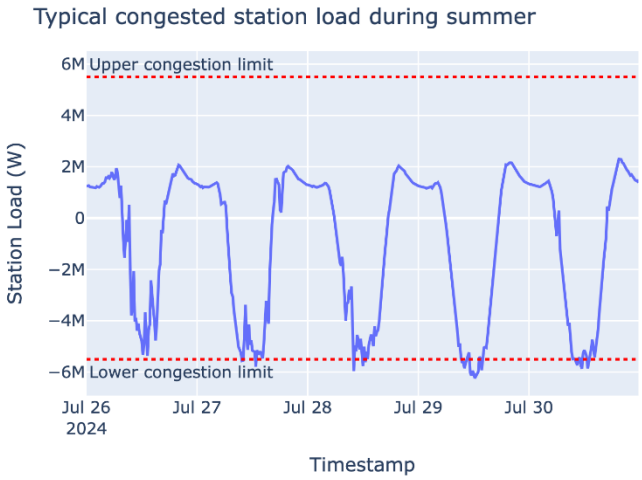

Figure 1: Typical power load pattern at a grid node whereby future loads and in particular loads surpassing safety limits (‘congestion’) need to be forecasted.

As highlighted by the word ‘typically’, this does not regard a fixed algorithm for a precise task like, e.g., calculating the rate-of-change-of-frequency (RoCoF) from a voltage signal as to comply with a written standard [5], but rather with a dynamic, high-dimensional input space with many additional model parameters that might be changed. It is therefore impossible to test and validate the congestion forecasting method once and for all. Verification of the performance of a single model instance for a single prediction job with a single instance of input data is useful, but insufficient to generate enough trust in proper performance in the course of time. The framework therefore includes a variety of elements like monitoring the distribution of the input data that must be considered to ensure the congestion forecasting procedure is and remains trustworthy.

2.1 Input data quality

A very important aspect for reliable usage of data analytics is monitoring the input data quality. In the case of congestion forecasting, historically measured loads and weather forecasts are used. A sudden change in the accuracy of these inputs can degrade model performance. For instance, correcting a systematic measurement bias in the input data may lead to poorer model performance if the model is trained with biased data. So even an improvement in input data quality must be carefully handled.

2.2 Input data distribution

The similarity of the distributions of the input data and the data used for training the model must be assessed. Various simple and more complex statistical methods exist for this purpose and it is still a very active area of research [6]. For instance, if the weather forecasts predict higher windspeeds than contained in the training data, the load forecast will be unreliable. Will there be a very large amount of wind power production, or rather very small due to wind farms being shut down for safety reasons? The model may not be able to answer this question and therefore the distribution of the input data must be continuously monitored.

2.3 Curation of training data

For training an ML model, the training data must be properly prepared, be of constant, sufficiently high quality, and correspond to the situation of interest. All relevant scenarios must be present in the training data and the dataset must not be strongly biased to a particular scenario, which would pose the risk of adversely affecting the training process.

In congestion forecasting, the data must correspond to a fixed grid topology. When mitigation actions have been performed to prevent overloading of grid assets, the training data must be amended such that it reflects the situation that would have happened in the case no action would have been taken. This involves estimating the amount of power at the node of interest for the hypothetical situation in which the mitigation action would not have been performed. If this is not properly done, high power loads above the congestion limit may not occur anymore in the training data and, depending on the model type being employed, the congestion forecasting model may erroneously never predict overloads anymore, until they start happening in practice, possibly resulting in large damage to grid assets.

2.4 Model explainability

In recent years, deep neural networks have shown impressive performance in particular in the field of image processing and large-language-models. However, they also make stunning mistakes (‘hallucinations’), which seem to be directly related to the complexity of these models. Models that are more interpretable, like random forests or Gaussian process regression are less prone to such mispredictions.

As part of the proposed test and validation framework it is good practice to apply some explainability methods like LIME or SHAP from the ML community to assess if the model is sensitive to physically irrelevant data features [7]. In such a case, the model will be less trustworthy.

Classical sensitivity analysis, also called ‘local linear explanation’, based on varying a single input (or type of input, e.g. temperature time series) and assessing its effect on the prediction can also be considered. Depending on the complexity of the model, this may give valuable insights into the inner working of the model.

2.5 Measurand, metrics and scoring rules

In metrology, one of the first steps for performing a measurement is to unambiguously specify the quantity intended to be measured, shortly called the ‘measurand’. Similarly, for validating a data analysis method, the output to be validated must be specified. That may seem trivial, but in

practice it might take some thinking. In the case of congestion forecasting, an obvious candidate is the value ('point prediction') of the load at a grid node. A suitable metric for evaluating model performance could then be, e.g., the mean-absolute-error (MAE) of the predicted load versus the true load as measured afterwards (in case no mitigation actions were taken by the operator).

From the operator's perspective, the main result of the prediction job is an indication whether or not an action needs to be taken and, if so, at which time point and how large the predicted overload is, as to mitigate a significant risk on overloading grid assets. However, focusing on metrics that quantify the accuracy of predicted overloads alone can lead to preferring models that are only good in optimizing such metrics, while having poor performance for predicting power loads in general. This makes them less reliable from a physical perspective and therefore less trustworthy.

The forecasting model can also provide probabilistic forecasts, i.e., a probability distribution for the predicted load at a specific point in time (or even for a range of points in time), in practice given by some quantiles of the distribution (e.g., 2.5 %, 50 % and 97.5 % quantiles). To evaluate the performance of probabilistic forecasts, one can use frequentist success rates ('Is the true load 95 % of the times contained in the interval given by the 2.5 % to 97.5 % coverage interval based on the quantiles?') or so-called 'scorings rules' like Continuous Ranked Probability Score (CRPS) [8] that take into account the full predicted distribution and compare it with the true load.

Furthermore, the costs associated with failing to predict real congestion may be much higher (if a grid asset is ruined) than the costs incurred when erroneously predicting congestion and requiring a mitigation. So, cost functions reflecting financial costs involved with different model outcomes and resulting actions can also be highly relevant when judging model performance, which is also closely related to the operator's risk appetite [9].

In conclusion, selecting a sensible set of relevant measurands, metrics and scoring rules is an essential and non-trivial step for testing and validating grid data analytics in the most meaningful way, and the choices made will affect the results of performance assessment.

2.6 Detailed evaluation of historical performance

For any data analytic, it is essential to evaluate the performance on historical data for which ground true reference results are available. For some applications, like predicting major frequency deviations, historical event data is scarce, whereas for other applications, like congestion forecasting, there is a large amount of historical data that can be used to assess the performance of the method. In such a case it is worthwhile breaking down the historical performance into different subcategories, which leads to better understanding of the performance of the method.

For congestion forecasting, performance can be split into different types of nodes (e.g., nodes close to solar parks or wind parks), different months and seasons, different weather conditions, days of the week, time of the day, etcetera.

2.7 Uncertainty analysis

In contrast to assessing model performance based on metrics using historical data, in a metrology-inspired uncertainty evaluation the effect of the uncertainty of each individual input on the uncertainty of the output of the data analysis method is assessed. Input data uncertainty can be propagated through an already trained model using the procedure presented in the GUM [3] or in its first supplement which describes a Monte Carlo method [10]. To do this properly, the correlation of the uncertainties of single entries in the input data must be taken into account, both in the time series of a single variable as well as between different variables. The effect of data uncertainty on model training should be assessed as well.

In contrast to the metrological community, the ML community typically distinguishes between epistemic and aleatoric uncertainty [11], whereby the exact interpretation of these categories slightly differs amongst different authors. Epistemic uncertainty is often equated with 'model uncertainty' and is the uncertainty that can be reduced by having more data, which may also relate to obtaining a completely new type of input variable. More data can allow for a better model architecture as well as obtaining better values of the model parameters during model training. Complementary to this, aleatoric uncertainty is the irreducible uncertainty that cannot be reduced by obtaining additional data. It is similar to irreducible random measurement noise, and is sometimes called 'data uncertainty'. However, as one can generally never be fully sure that the apparent random behaviour in observed data (e.g., loads) cannot be partly explained by some new type of input variable (e.g., important sports events impacting the behaviour of many people), the distinction between epistemic and aleatoric uncertainty is somewhat artificial, although it can be pragmatic in specific contexts.

Whereas typical methods for uncertainty quantification from the ML community are data driven, for a metrological sound uncertainty evaluation, expert judgment including domain knowledge of the application is essential. All possible influence quantities on model performance must be considered, including quantities for which no measurement data are available.

For congestion forecasting, one may think of influence quantities affecting solar power generation like the expected amount of desert dust in rain or the amount of snow, in particular if there is no snow in the training data. Information with respect to grid maintenance that may lead to increased loads at other assets is also important. Specific changes in the grids (e.g., a new solar farm) as well as political and societal events (e.g., new regulations, an important sports match on television) can also affect model performance and such changes may be uncovered when assessing the distribution of the input data. As such, considering all possible sources of uncertainty is also related to monitoring the distribution of the input data as discussed in section 2.1, although the idea in this

section is to be aware of uncertainties in model predictions upfront, rather than only based on possibly encountering out-of-distribution data based on a statistical method.

A comparison and harmonization of methods and way of thinking as performed by the ML and the metrological community and applied to data-driven grid data analytics is part of the current research.

2.8 Worst-case performance and robustness

ML models and in particular intractable deep neural networks may return completely erroneous predictions for slightly modified input data, whereas the predictions were correct for the original input data. The smallest change in input values that can result in a different outcome, or the largest difference in outcome for a fixed size of the change, is a measure for the model's robustness and is called a 'counterfactual explanation' in the ML community. This is similar to assessing the worst-case performance of the model while modifying the inputs in line with their uncertainties. It is a valuable exercise to search for such worst-case performance inputs, as figures showing average model performance may mask such possibilities for incidental very poor performance.

2.9 Simulated reference data

The generation and usage of simulated reference data for training or validation purposes can have some significant advantages as they can be tuned to reflect any desired scenario, they can be generated in large quantities, and they possess a known ground truth reference value. A major risk of (purely) simulation-based data analysis methods is that such methods may perform much worse in practice than during the simulation studies, as the simulated data is usually a simplified version of what real data may look like.

For the study and validation of congestion forecasting methods, simulations may not be the first choice of method, as significant amounts of historical real data are usually available. Nevertheless, simulations can still contribute valuable insights. Whereas in real-life situations the true value of the power load distribution is not known, in a simulation it is and reference values for, e.g., all computed quantiles can be computed. It can be assessed, e.g., whether the trained model correctly recovers the true quantiles, how much training data is needed, and whether or not the results from uncertainty evaluation and explainability methods make sense. One can start with, e.g., a very simple sinusoidal model with some Gaussian noise and make the model gradually more complex by including a known statistical model for weather-based generation and consumption profiles derived from real data. A trustworthy congestion forecasting method must be able to perform well in such a statistically controlled environment, so this poses a benchmark for assessing its trustworthiness.

Simulation-based validation is not a guarantee for correct performance in real-life applications, but it still gives valuable insights into the performance of the data analysis method.

2.10 Software development process

Whereas ill-designed software may still have very good performance, good performance is nevertheless less likely when the design and development process are below state-of-the-art standards for software design. Software bugs may more easily occur and be less easy to solve in poorly structured software. Good software design practices like using a software versioning system like git [12], the registration of issues and automated testing highly contribute to trustworthy software.

3 Example results

Based on the draft framework presented in section 2, first steps have been undertaken for testing methods for congestion forecasting, in particular OpenSTEF [1]. To this purpose, Alliander has made available a dataset containing 55 load measurements of 5 different node types (solar and wind parks, transformers, substations and MV-feeders) as well as predictors including weather forecasts, usage profiles of large connection types, and EPEX day-ahead market prices [13]. At the moment of writing, the calculated results are not yet definitive and the results presented in this section are meant as an conceptual illustration for the working of the framework.

Full application of the proposed framework and the publication of extensive test reports for various grid data analytics is foreseen by the end of the GridData project in May 2028 and will be published on the project website [2].

In this example we focus on the evaluation of historical performance of two types of models split into various types of grid nodes. Such an analysis can actually directly be performed with the recently added Backtesting, Evaluation, Analysis and Metrics (BEAM) model in the OpenSTEF library. The models xgboost and gblinear were compared and results were plotted for a typical prediction task for five different node types and two different metrics, see Figure 2 and Figure 3. While both models perform approximately equal in terms of the relative MAE, which is a measure for the quality of the point prediction, the gblinear has a better performance as probabilistic forecast, as quantified by the relative CRPS.

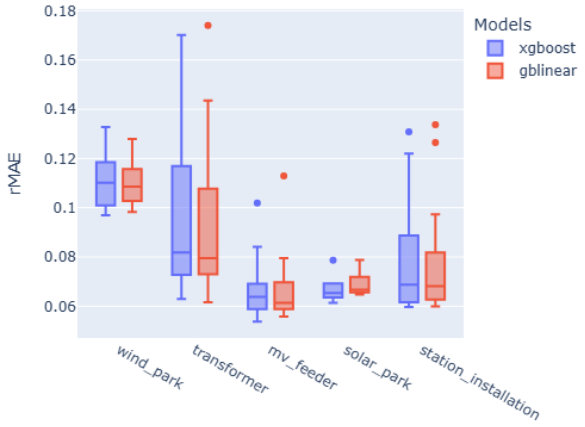

Figure 2: Results of historical performance testing of two different models for five different types of nodes. For comparability of the nodes, the MAE has been normalized by the observed range of loads at each location, yielding rMAE.

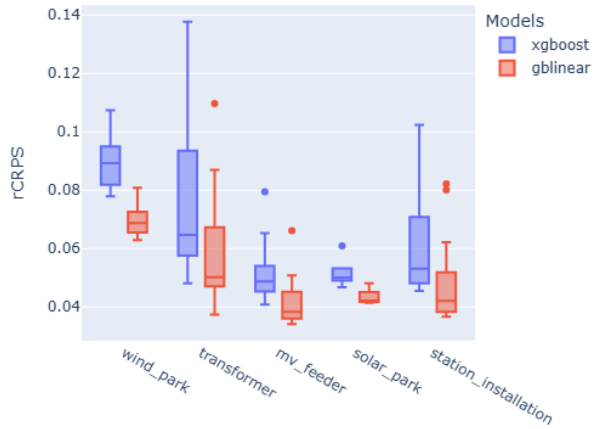

Figure 3: Similar to Figure 2, but this time the normalized CRPS has been used, i.e., rCRPS.

4 Conclusion

This paper presents a draft for a general framework for testing and validating grid data analytics as developed in the European metrology research project GridData. The framework is also meant to be applicable to methods based on ML. The European AI-act may require grid data analytics used for operating the grid to be validated by an external independent party. National Metrology Institutes like VSL can have an important role in this validation, in particular for applications with a relationship to measurements.

As ML models typically work in a dynamic context and can be retrained as well, a classical verification of the results computed by the software against some reference results based on a written standard will not be sufficient to assure the tested data analytic is trustworthy in practice. Instead, a set of validation methods is proposed.

Whereas the current work is performed in a research setting by the project partners, in a future setting the validating party may assess whether all aspects are properly covered by the party developing and/or using the software itself. Thoroughly testing and validating advanced ML-based grid data analytics using a standardized approach is essential for assuring continued reliability of grid data analytics and requires a joint effort from grid operators, software suppliers and independent knowledge institutes.

As the presented framework is still under development, the authors highly welcome feedback from the community.

5 Acknowledgements

The project (24DIT05 GridData) has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States.

References

[1] OpenSTEF, https://zenodo.org/records/18387491, accessed 23 February 2026
[2] 'Project 24DIT05 Metrology for Reliable Power Grid Data Analytics', https://griddata.inrim.it/, accessed 15 January 2026
[3] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008. doi:10.59161/JCGM100-2008E.
[4] Standard electricity user profiles, https://energiedatawijzer.nl/documenten/profielen-elektriciteit-2026/, accessed 23 February 2026
[5] IEC/IEEE 60255-118-1:2018: 'Measuring relays and protection equipment - Part 118-1: Synchrophasor for power systems – Measurements', 2018
[6] Lu, S., et al., 'Out-of-Distribution Detection: A Task-Oriented Survey of Recent Advances', ACM Computing Surveys 2025. https://doi.org/10.48550/arXiv.2409.11884
[7] Ali Öter, Betül Ersöz, 'Artificial Intelligence Assisted Solar Energy Forecasting by Explainability Approaches with LIME and SHAP', El-Cezeri Journal of Science and Engineering, Vol: 12, Iss: 2, 2025, pp. 205-212. https://doi.org/10.31202/ecjse.1591721
[8] Zamo, M., Naveau, P. Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. *Math Geosci* **50**, 209–234 (2018). https://doi.org/10.1007/s11004-017-9709-7
[9] Liang, F., et al., 'Risk–Cost Equilibrium for Grid Reinforcement Under High Renewable Penetration: A Bi-Level Optimization Framework with GAN-Driven Scenario Learning'. *Energies*, *18*(14), 3805. https://doi.org/10.3390/en18143805
[10] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method. Joint Committee for Guides in Metrology, JCGM 101:2008. doi:10.59161/JCGM101-2008.
[11] Hüllermeier, E., Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn* **110**, 457–506 (2021). https://doi.org/10.1007/s10994-021-05946-3
[12] Chacon, S. & Straub, B., 2014. *Pro git*, Apress. https://git-scm.com/, accessed 23 February 2026
[13] Liander 2024 Short Term Energy Forecasting Benchmark https://huggingface.co/datasets/OpenSTEF/liander2024-energy-forecasting-benchmark, accessed 23 February 2026