International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

POWER GRID STABILITY THROUGH AN AI AGENT FRAMEWORK: UNCERTAINTY-AWARE FORECASTING DURING EXTREME EVENTS

Praneesh Khanna[1]
University of Maryland, College Park, MD, praneesh.khanna1204@gmail.com

Abstract

Extreme weather events are increasingly destabilizing power systems by simultaneously damaging infrastructure and driving volatile demand. Traditional load-forecasting pipelines—designed for stationary dynamics and abundant data—are brittle in high-impact, low-frequency (HILF) regimes characterized by scarcity and heavy-tailed, multi-modal uncertainty. We propose TL-MDN, a two-stage AI agent that pairs transfer learning with a Mixture Density Network (MDN) head to produce calibrated, multi-modal probabilistic forecasts under rare-event conditions. A deep sequential encoder (e.g., LSTM/Transformer) is pre-trained on long-horizon "normal" operations from large markets (e.g., ERCOT, ISO-NE) and then adapted with an MDN head using sparse data from specific events (e.g., polar vortex, hurricane). We outline an evaluation protocol emphasizing proper scoring rules—CRPS, prediction-interval coverage probability (PICP), and Winkler score—to assess both reliability and sharpness. We further position TL-MDN against emerging LLM-based forecasters, highlighting complementary strengths and hybrid opportunities. The proposed framework targets deployment-grade usability for system operators through calibrated uncertainty, interpretable scenario modes, and seamless integration into reserve scheduling, demand response, and storage dispatch.

1. Introduction

1.1. Extreme weather and grid vulnerability

The electricity grid underpins modern society but is increasingly destabilized by climate-induced extreme weather. Hurricanes damage substations and lines, polar vortexes freeze pipelines, heatwaves cause equipment derating, and wildfires trigger preemptive shutoffs [1–4]. For instance, the February 2021 Texas polar vortex left millions without power due to fuel and generation failures [2]. These events illustrate the compounding vulnerabilities of aged infrastructure in a non-stationary risk landscape [5].

1.2. The forecasting dilemma: rarity and uncertainty

Accurate forecasts are critical for load balancing, reserve management, and outage prevention. Yet HILF events are rare, leaving historical datasets devoid of adequate training samples [6]. Classical machine learning models and deep learning pipelines—successful in abundant-data regimes—overfit or extrapolate poorly during anomalies [7]. Forecasting under these conditions requires not just point estimates but full probability distributions capturing multi-modal, heavy-tailed uncertainty [8].

1.3. The case for calibrated probabilistic forecasts

A point forecast conceals tail risks, leading to misinformed operator actions. In contrast,

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

probabilistic forecasts provide distributions over future outcomes. For grid control, calibration is essential: if a 95% confidence interval is predicted, the true value should fall within it approximately 95% of the time [9,10]. Proper scoring rules, such as CRPS, ensure models are evaluated on both sharpness and reliability [11].

2. A Critical Review of Forecasting Paradigms for Power Systems

2.1. Statistical and classical models

Traditional models like ARIMA capture seasonality and linear dependencies [12]. However, they assume stationarity and Gaussian residuals—conditions violated during abrupt, climate-driven disruptions. Hybrid approaches, such as ARIMA-LSTM, attempt to combine strengths but remain constrained by data limitations [13].

2.2. Deep learning and ensemble approaches

LSTMs and Transformers excel at nonlinear dependencies in abundant data settings [14]. However, under rare-event conditions, they overfit or collapse in extrapolation. Ensemble methods reduce variance by aggregating forecasts from multiple models, yet they remain bound by the limitations of the underlying data [15].

2.3. Tackling data scarcity

Researchers have explored augmentation (noise injection, time warping), synthetic generation (GANs, VAEs), and meta-learning strategies [16]. Still, these methods cannot invent wholly new event regimes. **Transfer learning**, by reusing representations from data-rich domains, emerges as a powerful alternative [17,18].

2.4. Probabilistic forecasting

Conventional uncertainty quantification assumes Gaussian errors. This underestimates extreme tail risks. Mixture Density Networks (MDNs) overcome this by predicting flexible, multi-modal distributions [19,20]. They are particularly well-suited to situations with multiple plausible load trajectories.

2.5. LLMs for time series

Large Language Models (LLMs) show promise in zero-shot forecasting by reframing numeric series as tokenized sequences [21]. They can also integrate unstructured side-information, such as storm advisories [22]. However, tokenization of continuous signals leads to noise sensitivity, and calibration remains poor [23,24]. Thus, LLMs are complementary but not substitutes for calibrated probabilistic models.

3. Proposed Framework: An Architecture for Rare Event Adaptation (TL-MDN)

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

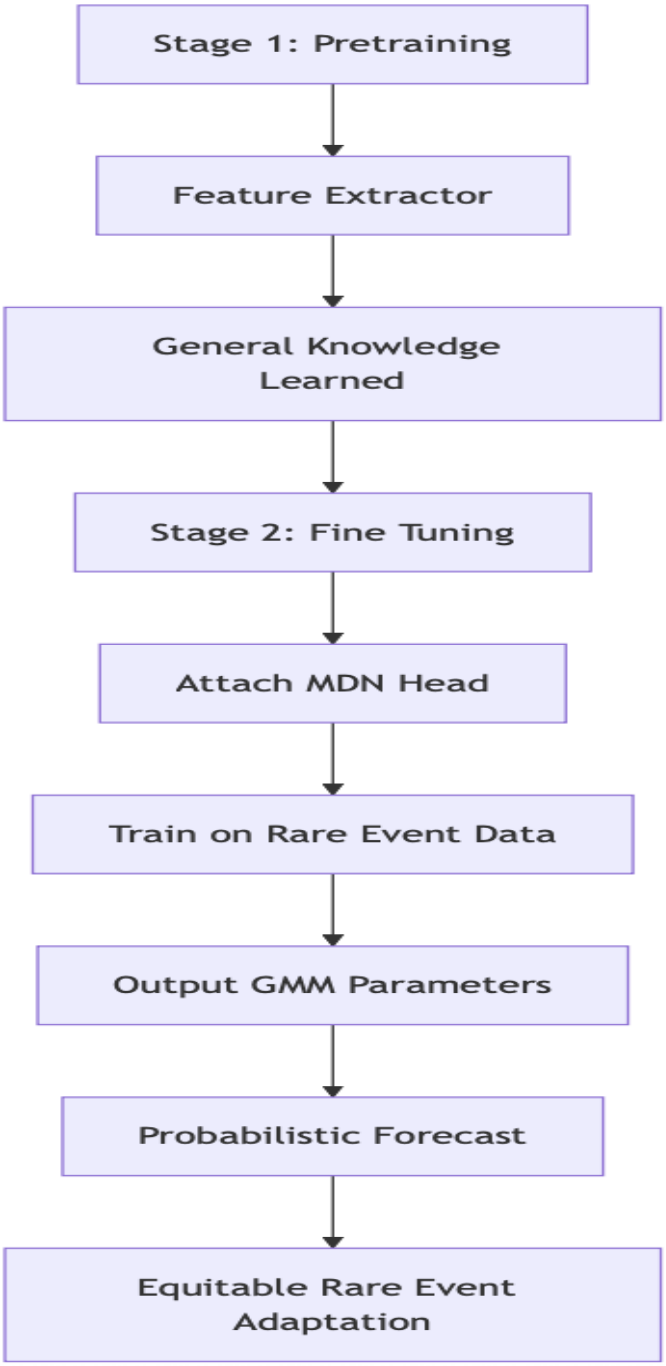

Figure 1: Proposed TL-MDN Framework for Rare Event Adaptation

3.1 Conceptual Design of the Forecasting Agent

The TL-MDN framework is conceptualized as an intelligent forecasting agent that can operate under conditions of extreme uncertainty. It is structured around two major components: a deep sequential encoder and a Mixture Density Network (MDN) head. The encoder, implemented using either an LSTM or Transformer architecture, is trained to ingest multivariate time series inputs such as historical load, temperature, humidity, and wind speed. During pre-training, the encoder is exposed to extensive datasets covering years of normal grid operations, enabling it to learn generalizable representations of load dynamics, including cyclical patterns and

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

nonlinear dependencies with weather variables [25,26]. This forms the backbone of the model's "normalcy knowledge."

The MDN head is then connected to this encoder and fine-tuned on sparse datasets collected during rare events such as the 2021 Texas polar vortex and Florida hurricane outages [2,4]. Unlike conventional regression heads that output a single predicted value, the MDN outputs the parameters of a Gaussian mixture model. This probabilistic output allows the model to represent multiple possible outcomes simultaneously, such as high demand spikes, collapses due to outages, or post-event recovery surges. Through its explicit learning of the probability distribution of outcomes, TL-MDN offers operators calibrated forecasts that reflect the underlying uncertainty of rare events [19,20].

3.2 Transfer Learning as a Mechanism for Rare-Event Adaptation

Transfer learning is the core to how TL-MDN can operate effectively in data scarce environments. Severe weather events by definition occur infrequently, and datasets recording their influence on power networks are small in number and variety. Deep neural networks, when trained only on such small datasets, overfit and generate predictions that fail to generalize to novel situations [6]. By pre-training the encoder on big data of normal grid operation, the model learns an excellent representation of the underlying dynamics of the grid. Fine-tuning thereafter only involves the acquisition of deviations from this baseline under conditions of extremes, significantly lowering the amount of event-specific data needed [17,18]. This method overcomes the "curse of rarity" through efficient reuse of knowledge from rich domains.

3.3 Probabilistic Representation Through Mixture Density Networks

The probabilistic forecasting of the TL-MDN is attributed to its MDN head. Rather than outputting deterministic values, the MDN outputs parameters of a mixture of Gaussian components with a mean, variance, and mixing weight. The model enables the forecast to be capable of recognizing skewed, heavy-tailed, or multi-modal distributions typical in the occurrence of rare events [19]. For instance, in the case of a hurricane, one would have a Gaussian one capturing the likelihood of stable demand, one would capture a collapse in the event of grid outages, and a third would capture post-event recovery surges. Their sum weighted by appropriate weights is an interpretable and flexible probability density function that is ideal for operational risk estimation [20]. Through offering probabilistic predictions scaled to the point scale in place of point predictions, TL-MDN is compliant with system operators' choice requirements for planning for ordinary outcomes and unforeseen hazards [9,10].

4. Experimental Design and Evaluation Protocol

4.1 Data Sources and Simulation of Rare Events

The experimental setup uses a two-stage data architecture to verify the TL-MDN model. The source pre-training data includes ERCOT and ISO-NE hourly load data from the years 2010–2020, along with meteorological data from NOAA including temperature, dew point, humidity, and wind speed [25,26].This dataset provides a broad and representative view of grid behavior under normal and moderately anomalous conditions. For the fine-tuning stage, smaller curated datasets are used, representing critical extreme events. These include the February 2021 ERCOT polar vortex, characterized by widespread load shedding and generator failures, and Florida hurricane datasets, which capture both storm-induced outages and subsequent recovery dynamics [2,4]. Together, these datasets simulate the transition from stable conditions to rare,

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

disruptive events.

4.2 Comparative Benchmark Models

To evaluate the performance of TL-MDN, its forecasts are compared with several benchmarks. A standard ARIMA model provides a statistical baseline, reflecting traditional approaches that assume linearity and stationarity [12]. Deep learning baselines include a standalone LSTM trained from scratch, which highlights the limitations of data-hungry models in scarce-data regimes, and an ensemble of LSTMs, which demonstrates how variance reduction through aggregation can improve robustness but often at the expense of sharpness [15]. An ablation study with an MDN trained from scratch isolates the benefits of transfer learning, while a Large Language Model (LLM) used in a zero-shot setting represents an emerging but unproven paradigm in time series forecasting [21,22].

4.3 Evaluation Metrics for Probabilistic Forecasts

The paradigm supports exact scoring rules for probabilistic forecasts. Continuous Ranked Probability Score (CRPS) is used in the computation of the quality of the complete predictive distribution by comparing cumulative distribution function forecast and observed values [10]. Prediction Interval Coverage Probability (PICP) captures calibration by checking that observed values fall within predicted intervals at the desired rate [9]. Winkler Score is a penalty metric that penalizes highly wide intervals and failure intervals to capture true values, thus satisfying sharpness as well as reliability [27]. Such metrics make tight and multiparameter analysis of the model.

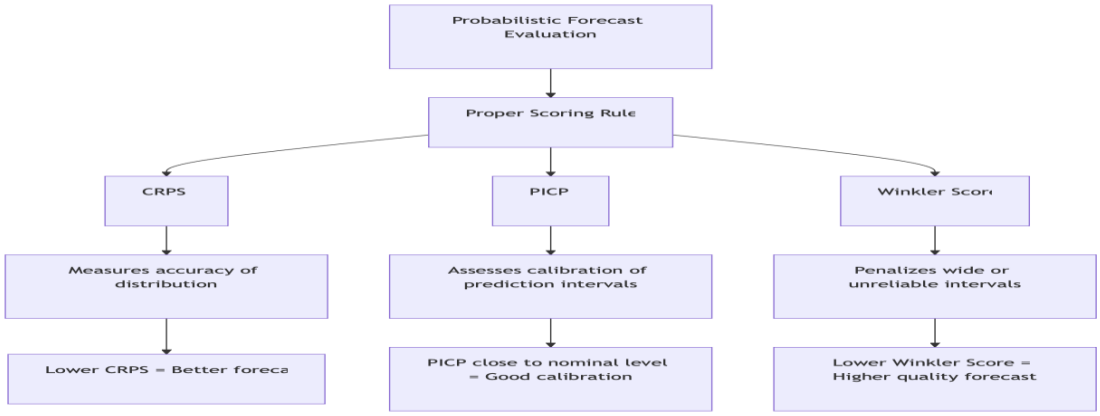

Figure 2: Metrics for Rigorous Probabilistic Forecast Evaluation

Table 1: Definition and Interpretation of Probabilistic Forecasting Metrics

Metric	What It Measures	Interpretation for Grid Operations	
CRPS	where F is the forecast CDF and y is the observation. [59]	**Overall Probabilistic Accuracy:** The integrated "distance" between the entire predicted probability	A consistently low CRPS indicates that the model's forecasts are accurate in both their central tendency (mean/median) and their spread (variance). It signifies a high degree

Received: August 01, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

		distribution and the single observed value. Generalizes Mean Absolute Error.	of overall forecast quality.
PICP	where [Li,Ui] is the prediction interval for observation yi. [62]	**Reliability / Calibration:** The empirical frequency that the true value falls within the predicted interval. Answers: "Is the model's stated confidence justified?"	A PICP value close to the nominal level (e.g., 0.95 for a 95% interval) means the operator can trust the model's uncertainty estimates for risk assessment. This is crucial for decisions like setting reserve margins or committing generation units.
Winkler Score	where α is the miscoverage rate (e.g., 0.05 for a 95% interval). [67]	**Balanced Sharpness & Reliability:** The width of the prediction interval, plus a penalty for observations that fall outside it. The penalty is proportional to the distance of the observation from the interval boundary.	A low Winkler Score indicates that the forecast provides prediction intervals that are not just reliable (high PICP) but also usefully narrow (sharp). This is essential for actionability; an interval that is too wide (e.g., "load will be between 50 GW and 150 GW") is reliable but useless.

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5. Anticipated Performance and Comparative Discussion

5.1. Expected Efficacy of the TL-MDN

The TL-MDN architecture should demonstrate clear superiority over normal statistical and conventional machine learning practices. From transfer learning, it is able to leverage the powerful inductive bias of a pre-trained encoder so that it can generalize extremely well from minute samples of infrequent events.

TL-MDN is therefore also expected to have lower Continuous Ranked Probability Scores (CRPS) compared to ARIMA, LSTM, and ensembling models, meaning better accuracy in the probability distribution estimation of outcomes [10,12]. In addition, its Prediction Interval Coverage Probability (PICP) levels are also expected to be closer to nominal confidence levels, meaning that the predicted intervals are well calibrated [9].

Deep models, even if they are retrained, suffer from overfitting, whereas ensembles, being more stable, also make the prediction interval too broad and thus lose their usability for operation [15].

5.2. Interpretability of Forecast Outputs

Another advantage that is highly referred to for the TL-MDN approach is the interpretability of its probabilistic outputs.

As opposed to black-box methods, which only give a point prediction, the MDN gives numerous Gaussian components with a likelihood to correspond to various physical events. To illustrate, in a hurricane scenario, one model might capture a flat demand scenario, another a collapse scenario caused by widespread outages, and another a recovery peak scenario as restoration occurs [19].

These types of interpretability enable operators to align model output with ground circumstances, and this fosters trust in autonomous forecasting systems. Conventional deep learning models, however, lack such ease of scenario-level reasoning and hence fall short of usefulness in critical operational conditions.

5.3. Comparative Analysis with LLM-Based Forecasting

Large Language Models (LLMs) were recently suggested as zero-shot or few-shot forecasters by re-framing time series as tokenized sequences [21,22].

These models are naturally suitable to handle textual side-information like weather warnings and news summaries, and providing more context to predictions [23]. However, their use to handle plain numeric time series is also restricted due to precision and calibration-related issues. Research has proven that LLM-based predictions are not stable, with small perturbations to inputs causing large differences in output [24].

Alternatively, TL-MDN is specifically built for calibrated probabilistic forecasting, featuring sharp distributions and accuracy. This implies that while LLMs can be ancillary tools for enhancing inputs to forecasting, TL-MDN is the wiser choice for large load forecasting under unusual events.

5.4 Limitations and Risks

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

While it is useful, the TL-MDN model is not flawless. One danger of negative transfer is: where source and target domains are very different from each other, information transferred from the encoder can mislead fine-tuning instead of helping it [17].

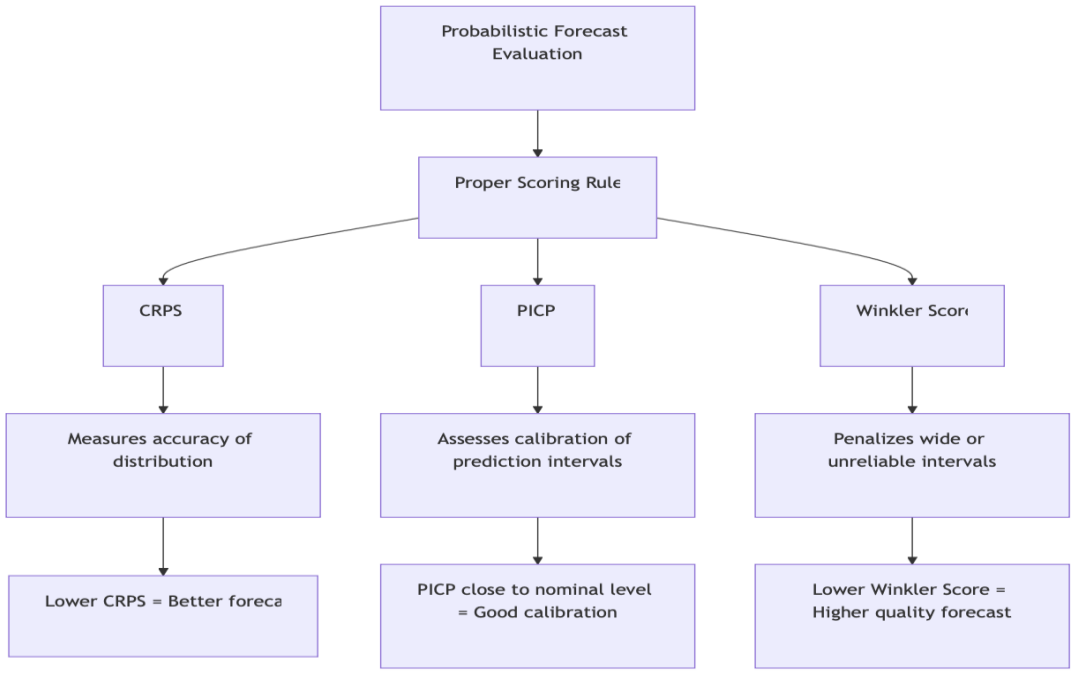

Figure 3:Comparative Strengths of TL-MDN and LLM-Based Forecasting

Table 2: Qualitative Framework Comparison: TL-MDN vs. LLM-based Approaches

Capability Dimension	TL-MDN Framework	LLM-based Framework
Data Efficiency on Target Task	**High:** Pre-training drastically reduces the data required for fine-tuning on the rare event.	**Medium to High:** Can perform zero-shot or few-shot forecasting, but performance improves with fine-tuning.
UQ Calibration & Reliability	**High:** Explicitly models a flexible probability distribution using an MDN, with a loss function that directly optimizes for likelihood.	**Low to Medium:** UQ is an emergent property of next-token prediction, not an explicit design feature. Deep learning models are often poorly calibrated without specific recalibration steps.
Handling of Unstructured Text Data	**Low:** Requires manual feature engineering to incorporate textual information as numerical	**High:** Native capability to process and reason over unstructured text, enabling fusion of multi-modal data.

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

	inputs.	
Robustness to Numerical Noise	**High:** Architectures like LSTMs and Transformers are designed to handle noisy, continuous time series data.	**Low:** The process of tokenizing continuous numerical values can amplify noise and lead to performance degradation.
Zero-Shot Capability	**None:** Requires at least a small amount of target data for fine-tuning to adapt to a new event type.	**High:** Core strength is the ability to perform tasks without specific training, by leveraging in-context learning.
Interpretability of Uncertainty	**Medium:** The components of the mixture model (means, variances) can sometimes be interpreted as representing distinct possible scenarios (e.g., "outage" vs. "no outage").	**Low:** The uncertainty is embedded in the softmax distribution over a vast token vocabulary, making it difficult to interpret in a physically meaningful way.
Computational Cost (Fine-tuning/Inference)	**Medium:** Fine-tuning is relatively fast. Inference is a single forward pass through a moderately sized network.	**High:** Fine-tuning large LLMs is computationally prohibitive for most organizations. Inference requires significant computational resources.

6. Implications for Real-World Grid Operations and Policy

6.1 Dynamic Reserve Management

The probabilistic outputs of TL-MDN enable a shift from static reserve planning to dynamic, risk-informed reserve management. Traditionally, operators rely on deterministic forecasts and maintain reserves based on fixed rules, such as the largest generator contingency. However, these approaches fail to account for the uncertainty and volatility of extreme events [2]. By using the full probability distribution of demand, operators can set reserves to cover the 95th or 99th percentile of expected load, ensuring that the system remains secure even in high-demand scenarios. This probabilistic approach balances reliability with economic efficiency, as reserves are only increased when forecasts justify the additional commitment.

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

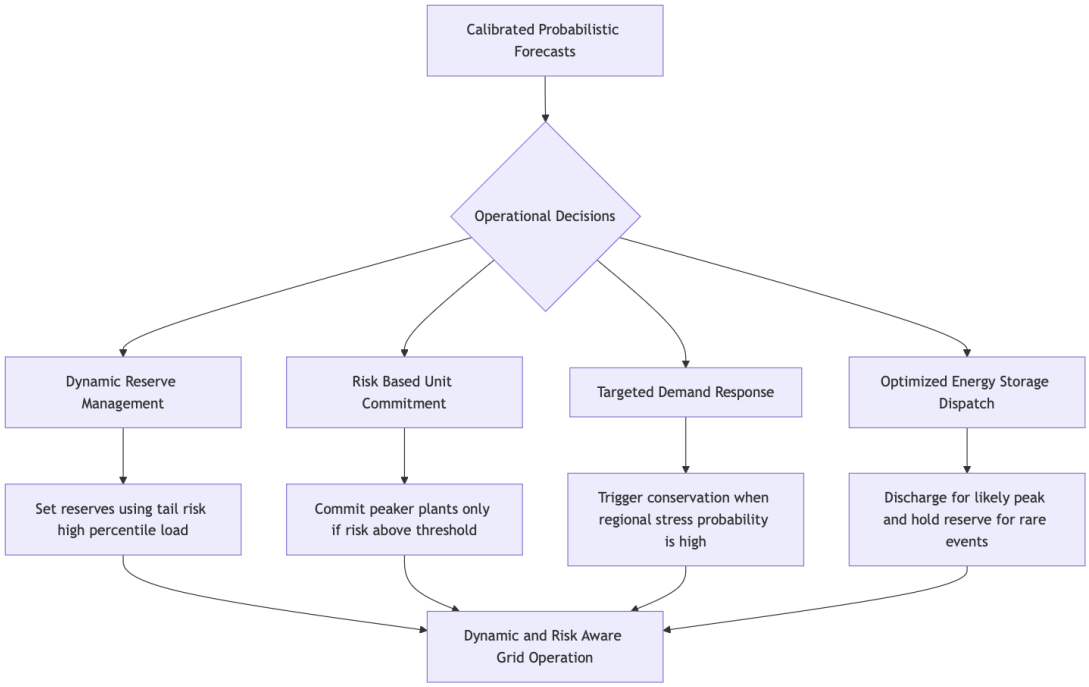

Figure 4: Integrating Probabilistic Forecasts into Grid Decision-Making

6.2 Targeted Demand Response Activation

Probabilistic forecasts also enrich demand response program design. Rather than across-the-board conservation appeals unreasonably annoying customers, operators can selectively apply demand response whenever the probability of system stress exceeds some threshold [4]. For example, in the event that there is a prediction of 30% probability of demand outstripping generation available, operators can apply selective demand curtailment in strained areas.

6.3 Optimized Energy Storage Dispatch

Large-scale batteries and other energy storage systems play a critical role in maintaining grid stability, especially during disasters. However, their effectiveness depends heavily on how well they are dispatched. Relying only on deterministic forecasts can lead to premature battery discharge, leaving the grid exposed to later demand spikes.

In contrast, probabilistic forecasts generated by TL-MDN provide operators with a more flexible framework to plan battery usage under different possible scenarios. This allows storage to be used for the most likely peaks in demand, while still reserving capacity to protect against less probable but high-impact events. This ability to hedge against risk is what makes storage assets valuable not just for efficiency, but also for resilience [19,20].

6.4 Policy and Regulatory Considerations

For probabilistic forecasting to reach its full potential in grid operations, policy and regulatory frameworks must evolve alongside it. Existing reliability standards are largely rooted in deterministic planning methods, which are increasingly inadequate in the face of climate-related uncertainty [5].

Received: August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Regulators will need to design standards that explicitly incorporate probabilistic risk analysis into operational planning. Clear reporting of forecast uncertainty will also be essential for building trust among stakeholders. Tools that visually communicate probability distributions, ranges, and scenario likelihoods can make these techniques more accessible for both operators and policymakers.

Future regulatory guidelines can act as a driver, pushing utilities toward advanced forecasting methods like TL-MDN, which better support planning for resilience.

7. Future Research

The emergence of TL-MDN marks a new era in power system forecasting, opening the door to a variety of exciting research avenues. One of the prominent avenues for going forward is to develop hybrid solutions that combine Large Language Models (LLMs) and TL-MDN.

Although TL-MDN provides numerically precise calibration in very high accuracy, LLMs are better at processing situational information like weather forecasts, news analysis, or social media trends. Put together, they can provide forecasts that are numerically accurate as well as infused with wider situational awareness.

Another significant avenue is physics-informed modeling, where fundamental principles—like generator limits and power-flow equations—are embedded directly into the learning process. This ensures that forecasts are not only statistically valid but also consistent with the physical realities of the grid.

The shift towards online or perpetual learning is another significant frontier. Rare as they are, "extreme events" provide rich information when they do happen. Models that can dynamically update themselves as responses will continuously improve their accuracy, better learn from shifting climate conditions, and advance resilience.

Finally, advances in multivariate probabilistic forecasting—simultaneously for renewable generation, load demand, and cross-regional power transfers—will allow operators to more accurately evaluate interconnected risks. This, in turn, provides support for more coordinated and secure strategies to ensure system stability.

8. Conclusion

Severe weather is quickly becoming among the most threatening risks to grid reliability. Conventional forecasting practices, based on steady conditions and heavy reliance on past history, are no longer adequate in a climate of frequent perturbations.

The TL-MDN model presented here overcomes these limitations in the first place by combining transfer learning with mixture density networks to generate calibrated probabilistic forecasts, especially of rare and extreme events. Through the creation of full probability distributions rather than point predictions, TL-MDN offers operators more in-depth insights for risk management, reserve planning, and early intervention.

A technical development at its core, TL-MDN also represents a larger transition—away from deterministic methods to probabilistic reasoning in grid operations. Such a perspective is essential in an age of uncertainty. Future advancements, such as the incorporation of LLM-based contextual information, physical system constraints, and continuous learning, will serve to further support this approach.

As the effects of climate change increase, predictive techniques such as TL-MDN—aiming to

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

specifically address uncertainty—will be key to the construction of resilient, flexible, and secure power systems.

References

1. Hawker, G., Ochoa, L., & colleagues. (2024). *Management of extreme weather impacts on electricity grids: An international review*. University of Strathclyde.
2. Energy Ventures Analysis. (2025). *Operation of the U.S. power grid during the January 2025 polar vortex*. America's Power.
3. National Renewable Energy Laboratory (NREL). (2022). *The evolving role of extreme weather events in the U.S. power system with high levels of variable renewable energy*. U.S. Department of Energy.
4. Chen, X., Zhao, X., Zhou, W., Zhang, Y., & Li, J. (2024). Climate change impacts on the extreme power shortage events of wind–solar supply systems worldwide during 1980–2022. *npj Climate and Atmospheric Science, 7*(1), 24–39.
5. Environmental and Energy Study Institute (EESI). (2022). *Living with climate change: The polar vortex*[Briefing]. Washington, DC.
6. Wang, H., Liu, Y., & Zhang, J. (2023). A comprehensive survey on rare event prediction. *arXiv preprint arXiv:2309.11356*.
7. Abolghasemi, M., Fasiolo, M., & Sottile, G. (2023). A machine learning model ensemble for mixed power load forecasting across multiple time horizons. *Scientific Reports, 13*, 10304500.
8. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.
9. Ziel, F., & Berk, K. (2021). An introduction to multivariate probabilistic forecast evaluation. *Patterns, 2*(1), 100012.
10. Lerch, S. (2017). *Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts* (Doctoral dissertation). University of Heidelberg.
11. Bishop, C. M. (1994). Mixture density networks. *Neural Computing Research Group, Aston University*.
12. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.
13. Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews, 29*(5–6), 594–621.
14. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
15. Zhang, C., Zhao, Y., & Xu, Z. (2023). Ensemble methods for power load forecasting under uncertainty. *Frontiers in Energy Research, 11*, 944804.
16. Wen, Q., Sun, L., Yang, F., & Song, J. (2020). Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*.
17. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on*

Received: August 01, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Knowledge and Data Engineering, 22(10), 1345–1359.

18. Harutyunyan, H., Khachatrian, H., Kale, D., Steeg, G. V., & Galstyan, A. (2022). Transfer learning for clinical time series analysis using deep neural networks. *Patterns, 3*(4), 100500.
19. Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford University Press.
20. Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics, 3*(6), 422–440.
21. Kim, M., Lee, J., & Park, H. (2024). Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR 2024)*.
22. OpenReview. (2024). *Time-LLM: Time series forecasting by reprogramming large language models*. Retrieved from https://openreview.net
23. Rasheed, F., Shahid, M., Zhang, Y., & Wang, Y. (2024). Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*.
24. Xu, Z., Zhao, Y., & Zhang, C. (2025). Revisiting LLMs as zero-shot time-series forecasters: Small noise can break large models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*.
25. Electric Reliability Council of Texas (ERCOT). (2025). *Hourly load data archives*. Retrieved from https://www.ercot.com
26. ISO New England (ISO-NE). (2025). *Operational impact of extreme weather events: Final report*. Holyoke, MA.
27. MAPIE Contributors. (2023). *How to measure conformal prediction performance?* (Documentation v1.0.1). Retrieved from https://mapie.readthedocs.io

Received: August 01, 2025