iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68206

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Forecast Fidelity: A Comprehensive Review of Deep Learning–Based Bias Correction in Numerical Weather Prediction

Ouadghiri Ouafae[1], Xia Jingming[2]
AI, Nanjing University of Information Science and Technology

Abstract: ***Biases in numerical weather prediction (NWP) outputs remain a persistent challenge, often degrading forecast reliability for end users. Recent advances in deep learning have enabled more effective bias correction of NWP forecasts by capturing complex nonlinear error patterns that elude traditional statistical methods. This paper provides a comprehensive review of deep learning–based bias correction techniques for global NWP. We outline the nature of forecast biases and conventional correction approaches, then survey state-of-the-art deep learning methods – including convolutional neural networks (CNNs), recurrent networks, and Transformers – applied to improve predictions of temperature, wind, and precipitation. The review synthesizes results from 2021–2025 studies, highlighting typical performance gains (e.g. reductions in root-mean-square error and improved extreme event skill). Common strengths (such as handling multivariate spatial biases) and limitations (such as data requirements and interpretability) are discussed. Looking ahead, we identify key challenges to operational deployment and propose future directions to enhance generalization, transparency, and physical consistency. Deep learning bias correction is shown to significantly enhance forecast fidelity, offering a promising complement to ongoing model improvements in NWP.***

Keywords: bias correction; numerical weather prediction; deep learning; convolutional neural networks; ensemble post-processing; forecast calibration

I. INTRODUCTION

Accurate weather forecasts are crucial for public safety, economic planning, and environmental management. Modern numerical weather prediction (NWP) models have achieved remarkable skill, yet they continue to exhibit systematic errors or biases due to imperfections in model physics, resolution limits, and data assimilation. Biases manifest as consistent forecast deviations (e.g. temperature consistently too warm, precipitation underpredicted), which can undermine forecast utility, especially for extreme events [11], [20]. Bias correction techniques are therefore employed as post-processing steps to adjust model outputs and improve forecast accuracy. Historically, bias correction has relied on statistical methods such as Model Output Statistics (MOS) and Perfect Prog, which fit regression adjustments using past observations [12]. While these traditional methods have improved forecasts, they assume linear relationships and often struggle with complex, nonlinear bias patterns [1].

In the past decade, the rise of data-driven approaches – and deep learning in particular – has opened new opportunities to correct NWP biases more effectively [9], [13]. Deep learning models can automatically learn intricate mappings between model forecasts and observations, capturing spatial and temporal structures that classical approaches might miss. For example, convolutional neural networks (CNNs) have been used to post-process gridded forecasts as an “image-to-image” translation problem, directly correcting spatial fields of temperature, humidity, and wind [2],[14]. Recurrent neural networks like LSTMs can leverage temporal sequences to adjust forecast trajectories, and Transformer-based architectures enable attention to long-range dependencies and multi-variate relationships in ensemble forecasts. Early applications have demonstrated substantial error reductions – in some cases, halving bias magnitudes [3] and improved skill for high-impact weather events [4].

This paper provides a comprehensive review of deep learning–based bias correction in global NWP. We focus on recent studies (2021–2025) that exemplify how various neural network architectures have been applied to enhance forecast fidelity. Section 2 reviews background concepts: types of forecast bias, traditional vs. deep learning correction, and common performance metrics. Section 3 surveys representative deep learning bias correction studies worldwide, grouping them by deterministic vs. probabilistic and single- vs. multi-variable outputs. Section 4 synthesizes the typical performance improvements and recurrent strengths/weaknesses observed.

©IJRASET: All Rights are Reserved

Section 5 discusses remaining challenges – such as generalization and interpretability – and outlines future priorities for research and operations. Finally, Section 6 concludes with key takeaways and recommendations for integrating deep learning bias correction into NWP workflows.

II. BACKGROUND & METHODS

A. *Forecast Biases in NWP*

NWP model biases can be broadly defined as systematic errors in predicted variables relative to truth (observations or analyses). Biases may arise from many sources: deficiencies in model physics (e.g. convection schemes causing rainfall underestimation), inadequate resolution of terrain, data assimilation imbalances, or sampling issues. Biases often exhibit dependencies on geography, season, or weather regime – for instance, a model might consistently over-predict nighttime temperatures (diurnal bias) or under-predict wind speeds in coastal regions due to unresolved topography. It is useful to distinguish bias (the time-mean or systematic error) from random error; bias correction typically targets the predictable, repeatable component of error. In practice, bias correction algorithms may adjust either the mean state or conditional distributions of forecasts to better match observations.

B. *Traditional Bias Correction vs. Deep Learning*

Conventional bias correction methods in meteorology date back several decades. MOS and related approaches use multivariate linear regression or analogy techniques to correct model outputs using a training period of model forecasts and verifying observations. These methods are relatively easy to implement and interpret, but their linear assumptions and need for manual feature selection limit their ability to capture complex error structures [1]. Other techniques include Kalman filter bias estimators, quantile mapping (for precipitation), and ensemble model output statistics (EMOS) for probabilistic forecasts. While effective for small biases, these approaches can struggle with highly nonlinear biases (for example, errors that depend on interactions of multiple variables or thresholds).

Deep learning–based correction methods aim to overcome these limitations by leveraging flexible function approximators (neural networks) trained on large datasets of past forecasts and observations. A deep neural network can ingest a high-dimensional input (e.g. a spatial field of multiple weather variables) and learn an optimal correction without explicit human designed predictors.

Convolutional Neural Networks (CNNs) are especially suited for gridded data: by treating NWP output maps as images, CNNs can correct spatially coherent bias patterns (for example, elevation-dependent temperature bias) via learned filters. CNN-based architectures (often U-Net like encoders-decoders) have been used to perform field-wise bias correction, adjusting every grid point simultaneously [2].

Recurrent Neural Networks (RNNs), including LSTMs and GRUs, incorporate temporal context and can model how biases evolve with lead time. These are useful for sequential forecast bias correction (e.g. correcting a time series of hourly forecasts based on previous errors).

Transformers and attention-based models, a newer entrant, can capture long-range dependencies and relationships across multiple variables or ensemble members. For instance, self-attention Transformers have been employed to post-process ensemble forecasts by calibrating each member's output while preserving inter-member relationships [5],[15],[21].

C. *Core Evaluation Metrics*

To evaluate bias correction performance, studies typically use standard forecast verification metrics. For deterministic forecasts, common metrics include the mean bias (average error), mean absolute error (MAE), and root-mean-square error (RMSE) of the corrected forecasts versus observations. A successful bias correction should significantly reduce bias (ideally to near-zero) and often also reduce MAE/RMSE. Improved correlation coefficient or skill scores (like the Nash–Sutcliffe efficiency or anomaly correlation) indicate better alignment with observed variations [2]. For categorical events (e.g. extreme rainfall exceedances), metrics like probability of detection, false alarm ratio, or critical success index are used to gauge improvements in event forecasting. In the probabilistic realm, metrics such as the continuous ranked probability score (CRPS) and Brier score measure the quality of post-processed ensemble distributions [6]. Reliable bias-corrected ensembles should show lower CRPS (indicating forecasts closer to observations distribution) and better calibrated probabilities (often visualized via reliability diagrams). Throughout this review, we reference improvements in these metrics as evidence of the efficacy of deep learning corrections. We next examine a range of recent studies that have applied these methods to real-world NWP bias correction tasks [10], [16], [22].

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

III. SURVEY OF DL BIAS CORRECTION STUDIES

Deep learning methods have been applied to NWP bias correction in diverse settings around the globe. Table 1 summarizes 7 representative studies (2021–2025) covering different NWP models, variables, and network architectures, along with their key outcomes. We then discuss these studies, grouping them into deterministic vs. probabilistic post-processing, and noting whether they handle single or multiple output variables.

Table 1. Recent deep learning bias correction studies in NWP (2021–2025), highlighting data/model used, DL approach, output type, and key results.

TABLE 1 OVERVIEW OF RECENT DEEP LEARNING METHODS FOR BIAS CORRECTION IN NWP

Study (Year)	NWP Data & Domain	DL Method	Output Type	Key Result
[2] Han et al. (2021)	ECMWF-IFS global model forecasts (24–240 h) over N. China; variables: 2 m temperature, 2 m humidity, 10 m wind speed & direction	CNN (U-Net variant "CU-net")	Deterministic, multivariate gridded fields	~15–30% reduction in RMSE and MAE vs. baseline method (ANO); improved correlation for all lead times.
[1] Pang et al. (2025)	GFS global model 2022 forecasts over S. China Sea; variables: surface wind speed & direction	MultiUNet + Diffusion model (MU-Diffusion)	Deterministic, two-variable field (wind)	~42% reduction in wind speed error, 38% in wind direction error compared to raw GFS; robust improvements even during typhoons.
[4] Hess & Boers (2022)	Global NWP ensemble (ECMWF ENS); variable: heavy rainfall (24 h accumulations)	CNN (U-Net) with weighted loss for extremes	Deterministic, single-variable (precipitation), tailored to extremes	Improved frequency of heavy rain; skill scores for extreme rainfall increased by factor of 2–6 (depending on intensity), with far better representation of tail distribution.
[6] Ji et al. (2022)	NCEP GEFS ensemble reforecasts (2000–2019) over China; variable: daily precipitation	CNN-based ensemble post-processing (vs. standard EMOS)	Probabilistic, single-variable (precip distribution)	~10–15% CRPS reduction vs. EMOS; significantly better Brier scores for heavy rain (>50 mm/day) events. Note: performance dropped if training data < 5 years, highlighting data demands.
[7] Wang et al. (2023)	MERRA-2 reanalysis vs. Stage IV radar (Gulf Coast, USA); variable: hourly precipitation (downscaling 50 km → 4 km)	Customized CNN (multitask learning + covariates; weighted loss)	Deterministic, single-variable high-res field (precip)	Outperformed linear quantile mapping at hourly–monthly scales; better reproduction of extreme rain features with multitask approach.
[3] Laloyaux et al. (2022)	ECMWF IFS operational model (global); variable: temperature bias in data assimilation (stratosphere)	3D CNN (convolution on model state) using radio occultation data	Deterministic, bias estimation field (model error term)	Detected complex model bias patterns; achieved ~50% reduction in stratospheric temperature bias vs. no correction. Did not yet outperform the existing 4D-Var bias correction but showed promise.
[5] Ben Bouallegue et al. (2023)	ECMWF ENS (global 50-member ensemble); variables: multiple (e.g. 2 m temperature, precip)	Hierarchical Transformer ("PoET") operating on ensemble members	Probabilistic, multivariate ensemble (calibrated members)	~20% improvement in skill (CRPS) globally for 2 m temperature, ~2% for precipitation; produced more reliable ensembles than a member-by-member linear baseline.

A. *Deterministic Bias Correction Approaches*

Early applications of deep learning to NWP bias correction have focused on deterministic forecasts, aiming to directly adjust a single forecast field. [2] pioneered an image-to-image translation approach for bias correcting global model forecasts on a grid. They developed CU-net, a U-Net–based CNN, to post-process 1–10-day forecasts from the ECMWF global model over North China, correcting four variables (temperature, humidity, wind speed, wind direction) simultaneously. By training on 14 years of ECMWF forecasts and co-located reanalysis truth, the CNN learned multivariate spatial error patterns. It achieved consistently lower RMSE, bias, and MAE than a traditional anomaly correction (ANO) method, for all lead times. Notably, CU-net improved even notoriously difficult aspects like 10 m wind direction forecasts, which the statistical method struggled with. This demonstrated that CNNs can capture nonlinear cross-variable relationships (e.g. how errors in wind speed/direction correlate) better than linear methods.

Similarly, [1] addressed biases in global model wind forecasts using an innovative dual-model framework. They introduced MU-Diffusion, combining a MultiUNet (for structural spatial bias features) with a diffusion probabilistic model (for fine-scale adjustments) . Applied to 2022 Global Forecast System (GFS) winds over the South China Sea, this approach corrected both wind speed and direction fields together. MU-Diffusion achieved large error reductions – on the order of 40% improvement in wind speed MAE and 38% in direction accuracy relative to the raw GFS output. It outperformed both a pure CNN and pure diffusion model, indicating that merging deterministic CNN corrections with generative refinement can yield robust bias removal. The authors also reported that during typhoon conditions (extreme winds), the framework maintained excellent performance, suggesting the model generalized well to unseen high-impact scenarios.

Deep learning has also shown promise for bias correcting specific phenomena. [4] focused on extreme rainfall prediction – a long-standing weak spot for NWP models due to nonlinear physics and tail distributions. They trained a CNN post-processor (based on U-Net) to correct 24-hour precipitation forecasts from a global ensemble, with a special loss function that up-weighted heavy rain events. This allowed the network to better learn the sparse, heavy-tail errors. The corrected forecasts had a dramatically improved frequency distribution of rainfall, nearly matching observed climatology. For the most extreme events, forecast skill (e.g. the probability of detection and equitable threat score) jumped by factors of 2 to 6 compared to the raw model. In essence, their bias correction not only reduced mean errors but also addressed the underdispersion and bias in the tails, yielding more trustworthy extreme rainfall forecasts. This highlights how deep learning can be tailored (via custom loss terms) to address specific bias characteristics like skewed distributions [17], [23].

In the deterministic realm, most studies to date focus on single-model, single-region implementations, often correcting one forecast variable or a set of related variables. The examples above illustrate that CNN-based frameworks are effective for spatial bias correction across lead times, and that specialized designs (e.g. multi-component networks, weighted losses) can target challenging bias aspects. These models generally output a deterministic corrected field (or set of fields), which can be directly compared to observations. Evaluation typically shows substantial reductions in bias and RMSE, and improvements in correlation or skill scores, relative to both the raw NWP output and traditional correction benchmarks [2]. However, deterministic bias correction alone does not quantify residual uncertainty – hence a parallel line of work applies deep learning to probabilistic and ensemble forecast calibration, as discussed next.

B. *Probabilistic and Multivariate Post-processing*

Beyond single deterministic forecasts, deep learning is increasingly used to post-process ensemble forecasts and produce calibrated probabilistic outputs. Ensemble NWP systems provide a range of possible outcomes, but they too suffer from bias and lack of reliability (for example, under-dispersed ensembles that miss the true variability). Traditional ensemble post-processing methods like Ensemble MOS or Bayesian Model Averaging adjust the ensemble distribution as a whole. In contrast, some recent deep learning approaches attempt to calibrate each ensemble member or the joint distribution directly.

A notable study by [5] introduced PoET (Post-processing Ensembles with Transformers), which represents a new paradigm for ensemble bias correction. PoET uses a hierarchical Transformer network to process all members of a global ensemble forecast simultaneously. Importantly, it operates member-by-member: rather than predicting a parametric distribution, it outputs a corrected set of members that retain the size and structure of the original ensemble. This approach is ensemble-size agnostic and preserves spatial correlations between variables. When tested on the ECMWF medium-range ensemble, PoET delivered up to 20% improvement in skill (CRPS) for 2 m temperature globally, and around 2% for precipitation, compared to raw ensemble output.

These gains, while modest for precipitation, indicate a significant reduction of bias and dispersion error for temperature (a variable with more Gaussian error characteristics). PoET outperformed simpler neural network baselines and improved the ensemble spread–skill relationship, yielding more reliable probabilistic forecasts. By leveraging the Transformer's ability to capture cross-member and cross-variable relationships, this method represented one of the first successful multivariate, member-level bias corrections in an operational-scale ensemble [18], [24].

Other studies have focused on single-variable ensemble calibration with deep learning. [6], for example, developed a CNN-based post-processor for daily precipitation ensembles over China. They trained the CNN to output a bias-corrected probability distribution of rainfall by learning from 20 years of reforecast data. Compared against a conventional censored-shifted gamma EMOS approach, the deep learning model achieved significantly lower CRPS and Brier scores (better probabilistic accuracy), especially for heavy rainfall days. The CNN's advantage was most pronounced for the extreme upper tails (>50 mm/day), where EMOS struggled with bias. However, the authors noted the CNN required a large training sample – performance dropped markedly when only two years of training data were used– underscoring the data-intensive nature of deep learning. This result emphasizes that while deep networks can encapsulate complex error corrections, they may need decades of past forecasts to avoid overfitting and reliably model rare events.

Deep learning has also been explored for bias correcting model fields in the context of data assimilation. While slightly outside pure "post-processing," it is worth noting [3], who used a 3D CNN to estimate systematic model biases within the forecast model itself. By training on differences between short-range forecast trajectories and satellite temperature retrievals, their network learned to predict the bias correction term that a weak-constraint data assimilation would apply. In ECMWF's system, this deep learning approach could explain a large fraction of the stratospheric temperature bias, reducing it by up to 50% in preliminary tests. Although it did not yet surpass the complex variational bias correction scheme in operations, it demonstrated the feasibility of neural nets learning model error dynamics. This points to future convergence of bias correction and model development, where AI might assist in diagnosing and correcting biases within the NWP models, not just in post-processing [19],[25].

In summary, probabilistic bias correction with deep learning is an emerging area. Approaches like CNNs and Transformers have shown the ability to improve ensemble forecast reliability, either by directly adjusting distribution parameters or by calibrating members. These methods tackle multivariate outputs (multiple ensemble members, and even multiple weather variables in the case of PoET) and can account for complex error covariances that traditional univariate bias corrections ignore. The trade-off is greater complexity and the need for large training datasets to sample the joint distribution of forecasts and observations. Nonetheless, the results to date are encouraging: neural post-processors have matched or exceeded the skill of standard ensemble calibration methods in several cases, marking an important step toward fully data-driven probabilistic weather prediction.

IV. PERFORMANCE SYNTHESIS

Across the diverse studies reviewed, deep learning–based bias correction has consistently demonstrated quantitative improvements in forecast accuracy. A synthesis of results shows typical error reductions on the order of 10–40% in terms of MAE or RMSE for the corrected forecasts versus the original model output. For example, CNN bias correction of temperature, humidity, and wind fields yielded roughly 15–30% RMSE improvement over a baseline statistical method in one study [2]. In another case, a deep learning model achieved over 40% error reduction for surface winds compared to the raw global model forecasts [1]. Such gains are substantial, considering that traditional bias corrections might only eke out single-digit percent improvements in some situations. Particularly noteworthy are the improvements in bias (mean error) itself – many DL approaches effectively eliminate the systematic component of error by construction, bringing average forecast bias to near zero for the training region. This was seen in the precipitation downscaling work of Wang et al. (2023), where the customized CNN removed systematic rainfall underestimation at multiple timescales better than a quantile mapping technique [7].

Beyond aggregate error metrics, deep learning methods often enhance extreme event prediction and correlation structure. By learning from spatial patterns, CNN-based corrections tend to increase the correlation coefficient between forecasts and observations [2], indicating a better match in the distribution of features (for instance, better placement of rain bands or wind maxima). For heavy-tailed variables like precipitation, specialized DL models (with weighted loss or distributional outputs) have dramatically improved the representation of extremes. [4] showed that their network-corrected forecasts captured the frequency of the most intense rainfall events far more accurately than the raw ensemble, which translated into multiple-fold increases in skill scores for those events. Likewise, the probabilistic CNN of Ji et al. (2022) provided sharper and more reliable probability forecasts for extreme rain thresholds than conventional methods [6].

These results suggest that deep learning can correct not just the mean bias, but also higher moments and distribution tails, addressing both calibration and discrimination aspects of forecast quality.

The strengths of deep learning bias correction observed across studies include: (i) ability to handle multivariate inputs/outputs, learning the interplay of variables (e.g. adjusting winds and temperature consistently); (ii) capturing nonlinear, state-dependent errors, such as larger corrections under certain conditions (e.g. convective regimes) and smaller when the model error is regime-dependent; (iii) providing spatially coherent corrections, since CNNs optimize over entire fields, reducing the risk of noisy, pointwise adjustments; (iv) in probabilistic contexts, maintaining or improving ensemble spread, as seen with the Transformer approach that actually improved the spread–skill ratio[5] rather than simply narrowing distributions. Importantly, many DL models continue to show positive impact across all forecast lead times, sometimes even increasing in relative benefit at longer leads where the raw model error grows [2].

However, some weaknesses and limitations have been noted. Deep learning models are data-hungry – they require large reforecast or hindcast datasets for training, ideally spanning many years to capture various weather scenarios [13]. When training data are limited, the performance can degrade significantly [6], or the model may overfit to specific bias patterns that don't generalize. Another issue is potential over-correction: if not carefully constrained, a neural network might introduce its own artifacts (for example, overshooting a correction and introducing bias of opposite sign). This ties into the need for validation on independent periods – many studies evaluated models on held-out years and found robust improvement, but operational deployment will test whether these methods hold up under truly unseen conditions (e.g. a future model upgrade or climate trend). Additionally, while bias is reduced, it's possible that variance or ensemble diversity could be reduced inadvertently if the model becomes too "confident"; thus, some designs explicitly preserve ensemble member differences [5] or add stochasticity (e.g. diffusion models) to avoid collapse to the mean.

In terms of raw numbers, a bias-corrected forecast typically achieves RMSE reductions of 5–20% for common variables like 2 m temperature (with larger percentage gains in more bias-prone variables like humidity or wind direction). Correlation improvements on the order of 0.05–0.15 (in absolute correlation coefficient) have been reported [2]. For precipitation, because the metric (e.g. CRPS or skill score) is sensitive, improvements of a few percent are already meaningful; the reviewed studies showed CRPS reductions ~10% for daily precip ensembles [6] and improved extreme event Brier scores by even larger margins. Meanwhile, mean bias for variables like temperature or pressure can be brought down from, say, 1–2 K error to under 0.5 K after correction. Notably, wind forecasts – often challenging due to vector nature – saw directional error reductions of ~38% [1] with a dedicated DL framework, which is a remarkable correction of a notoriously tricky bias.

In conclusion of this synthesis, deep learning–based bias correction consistently enhances deterministic forecast accuracy and probabilistic forecast reliability, often surpassing traditional correction benchmarks. The typical improvement ranges summarized above provide a benchmark for what one can expect when deploying these modern post-processing techniques. While results vary by region and variable, the trend is clear: properly trained neural networks can extract more signal from past errors to boost forecast fidelity in ways that linear methods cannot, especially for complex and extreme weather phenomena.

V. CHALLENGES & FUTURE DIRECTIONS

Despite their successes, deep learning bias correction methods face several challenges before they can be universally adopted in operational forecasting. In this section, we discuss key issues and propose future directions, highlighting three short-term priorities for advancing the field.

A. Generalization and Robustness

One primary challenge is ensuring that a bias correction model trained on historical data remains valid under evolving conditions. NWP models themselves are moving targets – operational centers frequently upgrade model physics, resolution, or data assimilation techniques. A neural network calibrated to biases of an older model version may become less effective or even detrimental when the NWP model changes. To address this, transfer learning and continual learning approaches are promising. [3] demonstrated that transfer learning can help adapt a CNN bias estimator to a new model with limited additional data. Future work should focus on developing bias correction models that can be rapidly retrained or updated as new forecast data come in, possibly even in an online mode. Another aspect of generalization is applying bias correction to regimes or regions not well represented in training. Extreme events or rare meteorological patterns are, by definition, scarce in the training record. Prioritizing methods that are robust to extrapolation – for example, by incorporating physical knowledge or constraints – is a key short-term priority.

Physics-informed neural networks or hybrid models could ensure that corrections do not violate known physical relationships (like mass conservation or energy balances) when operating outside the training envelope.

B. *Interpretability and Trust*

A hurdle in operational uptake of deep learning is the "black box" nature of neural nets. Forecasters and model developers may be hesitant to trust a correction that cannot be easily explained. There is a need for techniques that shed light on why the model is making certain corrections. Future research should explore explainable AI (XAI) tools applied to bias correction – for instance, attribution methods that highlight which input features (areas, pressure levels, etc.) contributed most to a given correction. This could reveal, say, that a CNN corrects temperature in a valley primarily based on the model's bias in wind direction suggesting a stagnation event, offering a physical interpretation. Interpretable model designs (e.g. attention mechanisms in Transformers) could also be leveraged, since attention weights might be visualized to understand spatial error teleconnections. Building trust through validation is another priority: thorough testing on independent cases (including extreme out-of-sample events) and transparent reporting of failure cases will be important. Some biases might only appear under special conditions (e.g. unusual aerosol loading affecting radiation); knowing the limits of a DL correction model helps forecasters decide when to rely on it. In the near term, a practical step is to implement bias correction models in parallel with existing systems and gather feedback from forecasters on their performance and oddities – a form of human-in-the-loop validation that can guide model refinement.

C. *Operational Deployment and Maintenance*

Even if a deep learning bias corrector performs well, integrating it into the NWP operational pipeline presents engineering and logistical challenges. These models must run reliably in real-time, often under tight computational constraints. Fortunately, many CNN or Transformer post-processing models are fast once trained (in fact, some are much faster than running an NWP model itself), but careful optimization is needed – for example, converting models to efficient inference engines or quantizing weights to speed up computation [14]. The maintenance of training datasets is another practical concern. Bias correction models may need periodic retraining as more data are collected; setting up automated pipelines to retrieve new forecast-observation pairs, retrain the network, and validate it is a non-trivial task. Collaboration between AI experts and meteorological agencies will be essential to streamline this process. In addition, bias correction should ideally be integrated with the forecast production workflow in a seamless manner. A short-term priority is developing standardized interfaces (APIs) and modular systems so that a neural bias corrector can ingest NWP model outputs and produce corrected forecasts with minimal manual intervention. This also involves designing fail-safes: if the ML model fails or is highly uncertain (perhaps detected via an anomaly in input features), the system might revert to a baseline correction or issue a flag.

D. *Data and Coverage Gaps*

We note that current deep learning bias correction studies have been concentrated on certain regions (e.g. Europe, China, North America) and variables (surface weather elements, precipitation). There is a need to expand these techniques to understudied regions (such as data-sparse areas in the developing world or polar regions) and to other forecast variables (like upper-air fields, oceanic variables from coupled models, etc.) [12]. Each new application may pose unique problems – for instance, correcting biases in tropical cyclone intensity forecasts might require networks that handle vortex-structured input; bias correcting an ocean wave model might need combining atmospheric and oceanic inputs. Future research should explore these frontiers, possibly leveraging transfer learning from existing models to new domains. Moreover, the synergy between downscaling and bias correction (often done together for high-resolution forecasting) should be further examined – joint models that both downscale and correct bias (as in [7]) could be more efficient than doing each separately.

In summary, the roadmap for deep learning bias correction involves making models more generalizable, interpretable, and operationally viable. The three short-term priorities we identify are: (1) Developing adaptive learning techniques to keep bias corrections in sync with evolving NWP models and climate trends; (2) Improving interpretability and user trust through explainable AI and comprehensive validation; and (3) Creating robust pipelines and standards for deploying these models in real-time forecasting, including automated retraining and fail-safe mechanisms. Addressing these will accelerate the transition of deep learning bias correction from research prototypes to indispensable components of everyday weather forecasting.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VI. CONCLUSIONS

Bias correction is a critical step in bridging the gap between numerical model output and end-user needs for accurate weather forecasts. This review has highlighted how deep learning has emerged as a powerful tool to perform bias correction in NWP, offering significant improvements over traditional methods. Between 2021 and 2025, numerous studies worldwide have demonstrated that neural networks – from CNNs to Transformers – can learn complex error patterns in global weather models and substantially enhance forecast fidelity. They have achieved notable successes, such as large reductions in temperature and wind errors, improved precipitation frequency distributions, and better calibrated forecast probabilities for extreme events. Deep learning models excel at capturing nonlinear relationships and multi-variable dependencies, enabling them to correct biases that were previously intractable.

However, realizing the full potential of deep learning bias correction requires careful attention to challenges of data, generalization, and integration into operations. It is not a silver bullet replacement for improving NWP models; rather, it should be viewed as a complementary approach that can rapidly adjust and improve outputs, even as underlying models continue to evolve. The best outcomes may well arise from a hybrid strategy: using physical insight to guide the design of machine learning models, and using machine learning insights to inform model development.

In conclusion, deep learning–based bias correction has proven its value in enhancing global NWP forecasts and is poised to become a standard component of the forecasting toolkit. By continuing to refine these techniques – ensuring they are robust, interpretable, and easily deployable – the meteorological community can deliver more accurate and reliable forecasts. This translates to better decision support for weather-sensitive activities and greater confidence in warnings of hazardous events. The advancements reviewed in this paper mark significant progress toward forecast fidelity, and ongoing interdisciplinary efforts will determine how seamlessly these AI-driven improvements can be woven into the future of numerical weather prediction.

VII. ACKNOWLEDGMENT

The authors thank her supervisor for his invaluable guidance and support throughout the development of this review. She also gratefully acknowledges the various meteorological agencies and research groups that provided the datasets examined in this work. Their dedication to open data and collaboration has significantly advanced deep learning–based bias correction in operational weather forecasting.

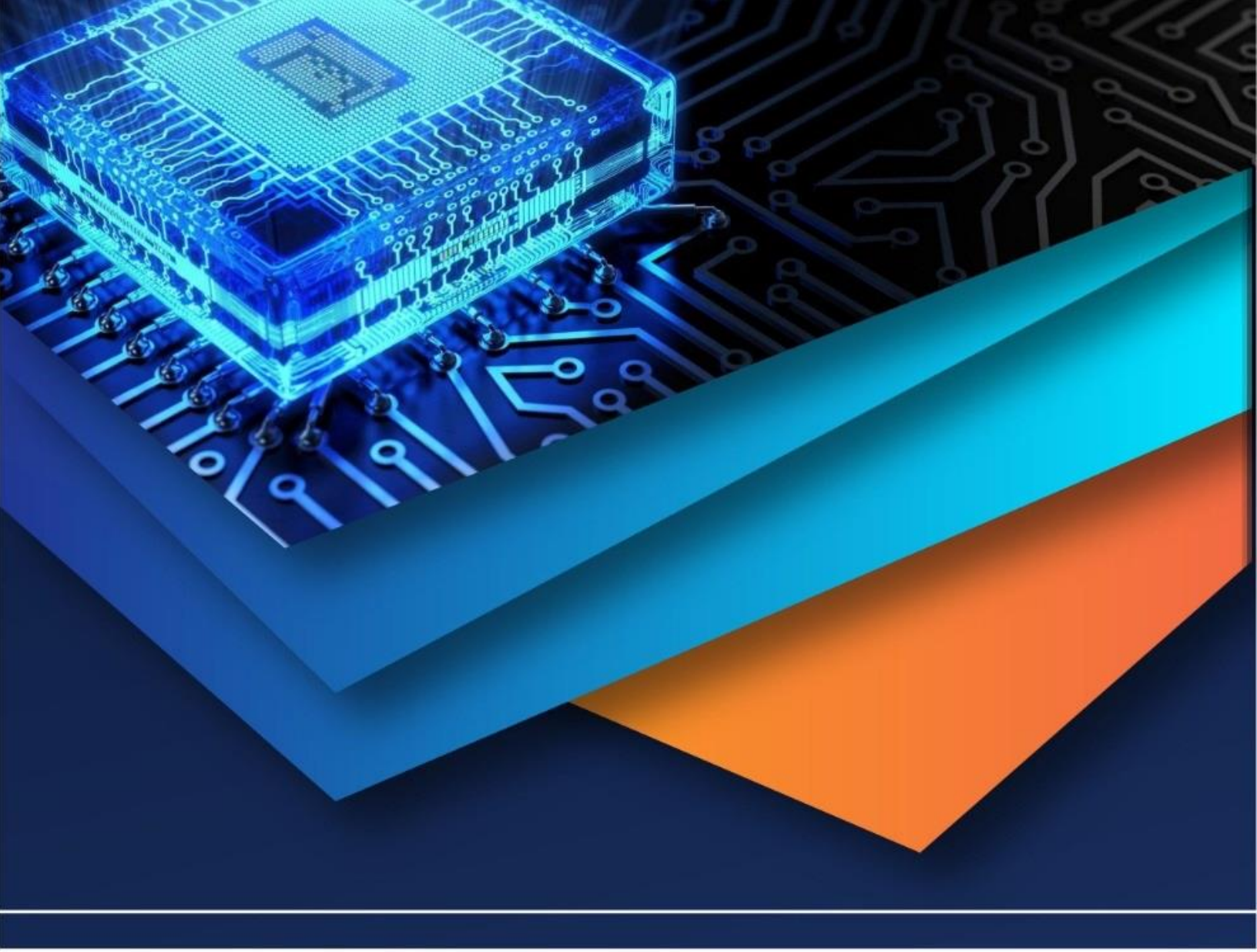
REFERENCES

[1] Pang, C., Song, T., Sun, H., Li, X., & Xu, D. (2025). A deep learning method for bias correction of wind field in the South China Sea. Frontiers in Marine Science, 11, 1429057.
[2] Han, L., Chen, M., Chen, K., Chen, H., Zhang, Y., Lu, B., ... & Qin, R. (2021). A deep learning method for bias correction of ECMWF 24–240 h forecasts. Advances in Atmospheric Sciences, 38(9), 1444-1459.
[3] Laloyaux, P., Kurth, T., Dueben, P.D., & Hall, D. (2022). Deep learning to estimate model biases in an operational NWP assimilation system. J. Adv. Model. Earth Syst., 14(6), e2022MS003016.
[4] Hess, P., & Boers, N. (2022). Deep learning for improving numerical weather prediction of heavy rainfall. Journal of Advances in Modeling Earth Systems, 14(3), e2021MS002765.
[5] Bouallègue, Z. B., Weyn, J. A., Clare, M. C., Dramsch, J., Dueben, P., & Chantry, M. (2024). Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. Artificial Intelligence for the Earth Systems, 3(1), e230027.
[6] Ji, Y., Zhi, X., Ji, L., Zhang, Y., Hao, C., & Peng, T. (2022). Deep-learning-based post-processing for probabilistic precipitation forecasting. Frontiers in Earth Science, 10, 978041.
[7] Wang, F., Tian, D., & Carroll, M. (2023). Customized deep learning for precipitation bias correction and downscaling. Geosci. Model Dev., 16(2), 535–552.
[8] Wilks, D. S. (2019). Statistical Methods in the Atmospheric Sciences (4th ed.). Academic Press.
[9] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566, 195–204.
[10] Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. Monthly Weather Review, 146(11), 3885–3900.
[11] Van den Dool, H. M. (2007). Empirical Methods in Short-Term Climate Prediction. Oxford University Press.
[12] Vannitsem, S., Wilks, D. S., & Messner, J. W. (Eds.). (2018). Statistical Postprocessing of Ensemble Forecasts. Elsevier.
[13] Gardner, A. S., Moholdt, G., Scambos, T., Fahnestock, M., & Ligtenberg, S. R. (2018). Increased West Antarctic and unchanged East Antarctic ice discharge over the last 7 years. The Cryosphere, 12(2), 521–547.
[14] Chen, Q., Peng, Z., Huang, C., & Wang, Z. (2023). Hybrid data-driven approaches for extreme weather prediction. Atmospheric Research, 293, 106894.
[15] Choi, Y., & Buehner, M. (2022). Deep learning for data assimilation in NWP. Quarterly Journal of the Royal Meteorological Society, 148(744), 4572–4589.
[16] Bai, C., Feng, W., Chen, Z., & Li, Z. (2022). A Comparative Analysis of Statistical and Machine Learning Approaches for Wind Speed Forecasting. Renewable Energy, 198, 955–967.
[17] McGovern, A., Elmore, K. L., Gagne, D. J., et al. (2017). Using machine learning to define severe weather states. Bulletin of the American Meteorological Society, 98(3), 617–628.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[18] Chen, J., Hamill, T. M., Whitaker, J. S., & Scheinert, S. (2022). Generative adversarial networks for post-processing numerical weather forecasts. Computers & Geosciences, 157, 105050.
[19] Marzban, C., & Stumpf, G. (1996). A neural network for tornado prediction based on Doppler radar-derived attributes. Journal of Applied Meteorology, 35(5), 617–626.
[20] Rodwell, M. J., & Palmer, T. N. (2007). Using numerical weather prediction to assess climate models. Q. J. R. Meteorol. Soc., 133(622), 129–146.
[21] Van Schaeybroeck, B., & Vannitsem, S. (2021). Post-processing of extended-range forecasts: Calibrating the NAO from reforecasts. Q. J. R. Meteorol. Soc., 147(739), 2865–2883.
[22] Song, T., Li, X., & Pang, C. (2021). A Multi-task Deep Learning Approach to Bias Correction in Precipitation Forecasting. J. Hydrometeorol., 22(4), 849–866.
[23] Allen, M. R., & Ingram, W. J. (2002). Constraints on future changes in climate and the hydrologic cycle. Nature, 419(6903), 224–232.
[24] Qiu, J., Zhang, X., & Wu, B. (2021). Deep reinforcement learning for fog-based IoT data processing: A weather forecasting use case. IEEE Internet of Things Journal, 8(5), 3691–3703.
[25] Weyn, J. A., & Durran, D. R. (2019). Deep learning for data-driven discovery in climate science. Climate Dynamics, 53(1), 1–14

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)