Received: 16 September 2025 | Revised: 22 December 2025 | Accepted: 6 January 2026

DOI: 10.1002/qj.70119

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

A composite-loss graph neural network for the multivariate post-processing of ensemble weather forecasts

Mária Lakatos

Faculty of Informatics, University of Debrecen, Debrecen, Hungary

Correspondence
Mária Lakatos, Faculty of Informatics, University of Debrecen, Kassai út 26, H-4028 Debrecen, Hungary.
Email: lakatos.maria@inf.unideb.hu

Funding information
Hungarian National Research, Development and Innovation Office, Grant/Award Number: K142849; National Research, Development, and Innovation Fund, Grant/Award Number: EKÖP-25-4-II

Abstract

Ensemble forecasting systems provide probabilistic estimates of future states, supporting applications from renewable energy production to transportation safety. Accurate forecasts are critical for operational decisions; however, systematic biases often persist, making statistical post-processing essential. Traditional parametric and machine-learning-based methods can produce calibrated predictive distributions at specific locations and lead times yet often struggle to capture dependencies across forecast dimensions. Multivariate post-processing methods, such as ensemble copula coupling and the Schaake shuffle, are therefore commonly applied to restore realistic inter-variable or spatio-temporal dependencies. This study applies a dual graph neural network (dualGNN) trained with a composite loss function that combines the energy score (ES) and the variogram score (VS) for the multivariate post-processing of ensemble forecasts. The method is evaluated on Weather Research & Forecasting (WRF)-based solar irradiance forecasts over northern Chile and European Centre for Medium-Range Weather Forecasts visibility forecasts for central Europe. The dualGNN consistently outperforms all empirical copula-based post-processed forecasts and networks trained only on continuous ranked probability score or ES, according to the multivariate verification metrics evaluated. For the WRF forecasts, its rank-order structure captures dependency information more effectively, improving the restoration of spatial relationships compared with the raw ensemble or historical observational ranks. Moreover, incorporating VS into the loss function also enhances univariate performance for both targets.

KEYWORDS

ensemble calibration, ensemble model output statistics, graph neural networks, multivariate post-processing, solar irradiance, visibility

1 | INTRODUCTION

Ensemble forecasts, generated by combining multiple model runs with varied initial conditions or configurations, provide a probabilistic framework that captures forecast uncertainty—an essential component for informed decision-making across a wide range of applications. Despite their advantages, these ensemble outputs often require calibration to correct for systematic biases (e.g., Hamill & Colucci, 1997). Statistical post-processing aims to produce calibrated forecasts that ensure consistency between observations and predictions, thereby enhancing probabilistic accuracy and overall reliability. This process is particularly important in regions where accurate environmental forecasts are critical for decision-making and the management of infrastructure.

A wide range of statistical post-processing methods has been developed to improve the calibration of ensemble forecasts; for a comprehensive overview, we refer the reader to Vannitsem *et al.* (2021). Among the most prominent are ensemble model output statistics (EMOS; Gneiting *et al.*, 2005) and Bayesian model averaging (Raftery *et al.*, 2005), both of which link ensemble forecasts to parametric predictive distributions and have been successfully applied to various meteorological variables. More recently, machine-learning approaches have received increasing attention in this domain. Distributional regression networks (Rasp & Lerch, 2018), for instance, directly map ensemble inputs to full predictive distributions and have demonstrated improved performance compared with traditional statistical methods across multiple applications. More advanced architectures—such as transformers, convolutional neural networks, generative adversarial networks, and distributional regression U-Nets—have subsequently been applied to post-process forecasts across a wide range of atmospheric variables (e.g., Bouallègue *et al.*, 2024; Dai & Hemri, 2021; Li *et al.*, 2022; Pic *et al.*, 2025).

A widely recognized limitation of most post-processing methods is that they are typically applied independently for each forecast horizon, location, and variable, potentially neglecting important dependencies across these dimensions. Multivariate post-processing aims to recover the dependence structure between different forecast dimensions, often lost in univariate calibration. This can be achieved using parametric approaches, such as Gaussian copulas (Möller *et al.*, 2013) or non-parametric copula-based methods, which reconstruct coherent joint predictive distributions without requiring a specific parametric form. These approaches are often termed two-step methods: first, samples are generated from the calibrated predictive distributions and then they are rearranged according to a dependence template. Examples include ensemble copula coupling (ECC; Schefzik *et al.*, 2013) and the Schaake shuffle (SSh; Clark *et al.*, 2004). Machine-learning-based approaches are increasingly applied in multivariate post-processing as well. In particular, generative models have shown promise in producing spatially coherent forecast scenarios, offering a flexible alternative to traditional copula-based techniques by directly learning complex dependency structures from data (Chen *et al.*, 2024).

Graph neural networks (GNNs) are an emerging approach in machine learning, yet their application to the post-processing of ensemble forecasts remains limited. A major strength of GNNs is their ability to model spatial dependencies, which is crucial for accurate meteorological forecasting. Feik *et al.* (2024) employed a GNN to generate calibrated forecast distributions for 2-m temperature predictions, and more recently Bülte *et al.* (2025) applied GNNs to improve extreme rainfall forecasts.

In this study, we employ a GNN to post-process ensemble forecasts, generating improved ensemble predictions directly at the output layer rather than estimating the parameters of the forecast distributions. Our focus is on solar irradiance and visibility, but the method is fully general and non-parametric, making it applicable to any meteorological variable and flexible with respect to the size of the post-processed ensemble. We investigate several variants of the GNN trained with different loss functions: one minimizing the continuous ranked probability score (CRPS), and another trained on the energy score (ES) following the approach of the conditional generative model by Chen *et al.* (2024). Our primary objective is to evaluate whether a GNN trained with a composite loss combining the ES and the variogram score (VS) can deliver improved predictive performance by simultaneously preserving overall distributional accuracy and better capturing spatial dependencies.

This article is organized as follows. Section 2 provides a brief description of the solar irradiance and visibility datasets studied. In Section 3 we review the univariate and multivariate post-processing methods applied, as well as the verification metrics used to evaluate forecast performance. Section 4 presents the details of the model implementations. The results of our two case studies are reported in Section 5, and this is followed by a brief discussion and concluding remarks in Section 6.

2 | DATA

The solar irradiance forecasts utilized here—also investigated by Baran *et al.* (2025)—were generated with the Weather Research & Forecasting (WRF) model (version 4.4.2; Skamarock *et al.*, 2019) for the year 2021. An

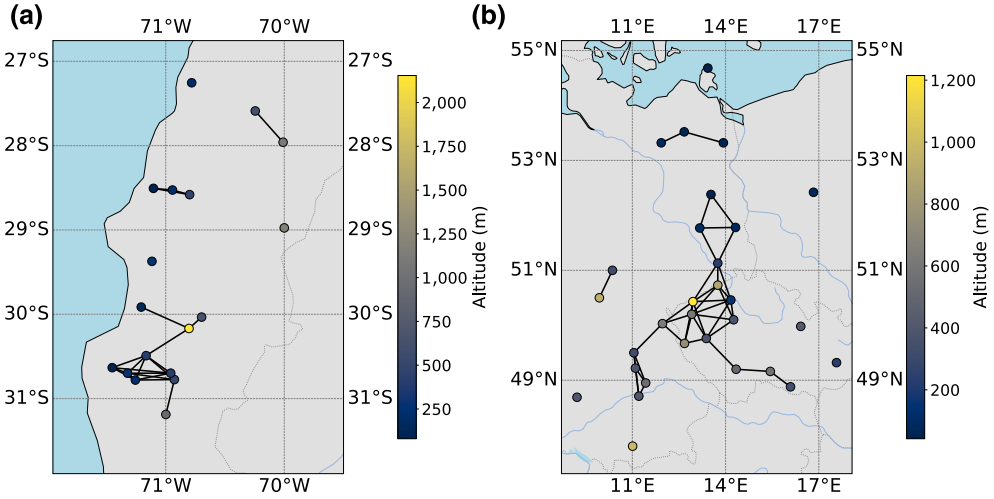

FIGURE 1 Locations of (a) solar irradiance observation stations in northern Chile and (b) visibility observation stations across central Europe. Black lines indicate edges used to construct the input graph for the GraphSAGE models. [Colour figure can be viewed at wileyonlinelibrary.com]

eight-member ensemble was employed, with a 3 km horizontal resolution and 36 vertical levels, capturing the key atmospheric layers. The model domain covers a wide range of Chilean landscapes, from the Pacific coast to the Andes. Observations, expressed in watts per square meter, were obtained from the National Weather Service's monitoring network (https://climatologia.meteochile.gob.cl/). After excluding stations with more than 10% missing data, 18 stations in the Atacama and Coquimbo regions (Figure 1a) remained for model training and evaluation. All forecasts were initialized at 0000 UTC, with a 1 hr temporal resolution and a forecast horizon of 48 hr.

The second case study examines visibility forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) at 30 SYNOP stations across Germany, the Czech Republic, and Poland (Figure 1b) for 2020–2021. Forecasts were initialized daily at 0000 UTC, with 20 lead times ranging from 6 hr to 120 hr at 6 hr intervals. Each ensemble includes the operational control run with unperturbed initial conditions, together with 50 additional members produced by the ECMWF Integrated Forecasting System, which are statistically indistinguishable, and hence exchangeable. Although ECMWF visibility forecasts are issued in 1 m increments and can therefore be treated as continuous, SYNOP stations typically report observations in discrete categories specified by World Meteorological Organization recommendations; namely, "100 to 5000 m in steps of 100 m, 6 to 30 km in steps of 1 km, and 35 to 70 km in steps of 5 km" (World Meteorological Organization, 2018, section 9.1.2). Accordingly, in Sections 5.2 and 5.4, the post-processed visibility forecasts are evaluated with respect to these 84 categories.

Note that, in the present study, for all models considered, only ensemble forecasts corresponding to the target variable were used as predictors, and forecasts of other meteorological variables were not included.

3 | METHODS FOR POST-PROCESSING AND FORECAST EVALUATION

This section details the univariate post-processing models applied to the solar irradiance and visibility ensemble forecasts. Section 3.1 describes the EMOS model used for solar irradiance forecasts, along with a multilayer perceptron (MLP) originally proposed by Baran *et al.* (2025). Section 3.2 describes the classifier employed for visibility forecast post-processing. Though Lakatos and Baran (2024) also applied an MLP to the raw forecasts, the proportional odds logistic regression (POLR) model generally achieved better performance and is therefore adopted here as the reference model.

3.1 | Univariate post-processing of solar irradiance

When modeling solar irradiance, it is necessary to account for its non-negativity and the frequent occurrence of zero observations; in such cases, probability distributions that assign a positive mass to zero are appropriate. Following the approaches of Baran *et al.* (2025) and Schulz *et al.* (2021), we employ the censored normal EMOS model, which enables a proper probabilistic representation of both positive and zero irradiance values.

For parameter estimation, we adopt the semi-local EMOS approach of Lerch and Baran (2017), which clusters stations with similar climatological and/or forecast error characteristics. Within each cluster, all stations share a common set of EMOS parameters, jointly estimated from the data of the stations in that cluster. This methodology facilitates more robust parameter estimation while preserving local representativeness, in contrast to the global

approach (Thorarinsdottir & Gneiting, 2010) that fits a single parameter set to data from all stations, potentially oversmoothing spatial variability.

Following the notation of Baran *et al.* (2025), let $G(x|\mu,\sigma) := \Phi((x-\mu)/\sigma)$, $x \in \mathbb{R}$, denote the cumulative distribution function (CDF) of a Gaussian distribution with mean μ and standard deviation $\sigma > 0$, where Φ is the standard normal CDF. The CDF of a Gaussian distribution with location μ and scale σ left-censored at zero is then

$$G_0^c(x|\mu,\sigma) := \begin{cases} G(x|\mu,\sigma), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

This distribution assigns a probability mass of $G_0^c(0|\mu,\sigma)$ at zero and has a mean given by

$$\kappa = \mu\Phi\left(\frac{\mu}{\sigma}\right) + \sigma\phi\left(\frac{\mu}{\sigma}\right),$$

where ϕ denotes the standard normal probability density function. Following Schulz *et al.* (2021), Baran and Baran (2024), and Baran *et al.* (2025), the ensemble statistics are linked to the parameters of the censored normal distribution through

$$\mu = \gamma_0 + \gamma_1\bar{f} + \gamma_2 p_0, \qquad \sigma = \exp(\delta_0 + \delta_1 \log S),$$

where γ_0, γ_1, γ_2, δ_0, $\delta_1 \in \mathbb{R}$ are model parameters, estimated following the optimum score principle of Gneiting and Raftery (2007), and p_0 and S^2 denote the proportion of zero observations and the ensemble variance respectively, defined as

$$p_0 := \frac{1}{K}\sum_{k=1}^{K}\mathbf{1}\{f_k = 0\}, \qquad S^2 := \frac{1}{K-1}\sum_{k=1}^{K}(f_k - \bar{f})^2,$$

where $\bar{f}$ is the ensemble mean and $\mathbf{1}\{\cdot\}$ denotes the indicator function.

An alternative to conventional statistical post-processing methods is the application of machine-learning models that directly produce calibrated ensemble forecasts. In this study, the MLP approach of Baran *et al.* (2025) is adopted, in which the number of output neurons is set equal to the target ensemble size. The network is trained by minimizing the sample CRPS, Equation (1), under the constraint that predicted solar irradiance values are non-negative. This distribution-free, data-driven framework allows flexible specification of the ensemble size, provided sufficient training data are available. To enable a fair comparison with the raw WRF forecasts, eight-member ensembles are generated for each prediction case. Previous experiments on the Chilean dataset have demonstrated that this approach outperforms the distributional regression network, which instead predicts the location μ and scale σ parameters of a left-censored normal distribution. In what follows, this method is referred to as MLP.

3.2 | Univariate post-processing of visibility

As mentioned in Section 1, visibility observations reported by the World Meteorological Organization form a discrete set of categories (World Meteorological Organization, 2018); as a result, the post-processing of visibility forecasts can be viewed as a classification problem. Following Lakatos and Baran (2024), the calibration of visibility ensemble forecasts is performed using POLR (McCullagh, 1980), a widely used parametric method for modeling ordinal response variables, which makes it particularly well suited for the visibility observations considered here, and thus serves as a powerful benchmark. The model specifies the conditional cumulative distribution function of the observed visibility Y given an M-dimensional feature vector $\mathbf{x}$ as

$$P(Y \leq y_i|\mathbf{x}) = \frac{\exp(L_i(\mathbf{x}))}{1+\exp(L_i(\mathbf{x}))},$$
$$L_i(\mathbf{x}) := \alpha_i + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta},$$
$$i = 1, 2, \ldots, 84,$$

where $\alpha_i \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^M$ are model parameters satisfying $\alpha_1 < \alpha_2 < \cdots < \alpha_{84}$. Consequently, fitting a POLR model to the visibility data requires estimating $84 + M$ parameters. In the present work, we apply the local POLR model, as also used by Lakatos and Baran (2024), for the calibration of visibility forecasts. In their study, this model provided the best predictive performance in terms of univariate performance across various spatial selection procedures, including semi-local and regional estimates (briefly discussed in Section 3.1), where local indicates that a separate model is fitted for each station based solely on its own historical data.

3.3 | Multivariate post-processing methods

To restore any potentially lost spatial dependencies, this study considers two-step approaches for multivariate post-processing as benchmark models. These benchmarks generate multivariate samples by combining forecasts that have been individually calibrated in a univariate manner using an empirical copula. Several variants exist depending on the strategy employed to define

the dependence template for restoring the dependencies. Extending the notation introduced in the previous section, let

$$\boldsymbol{f}^{(d)} = (f_1^{(d)}, f_2^{(d)}, \ldots, f_K^{(d)})$$

denote a K-member ensemble forecast for station d ($d = 1, 2, \ldots, D$) at a given time point and lead time.

The ECC (Schefzik *et al.*, 2013) leverages the rank-order structure of the raw ensemble forecasts, to retain the intricate dependency patterns present within the ensemble. Adopting the notation of Lakatos *et al.* (2023), the steps of this iterative procedure can be summarized as follows:

1. For each dimension d, generate a sample $\widehat{\boldsymbol{f}}^{(d)}$ of size K from the calibrated marginal predictive distribution, arranged in ascending order.
2. Define permutations

$$\boldsymbol{\pi}_d = (\pi_d(1), \pi_d(2), \ldots, \pi_d(K))$$

of the set $\{1, 2, \ldots, K\}$ corresponding to the rank order of the raw ensemble forecasts $\boldsymbol{f}^{(d)}$, where $\pi_d(k) := \text{rank}(f_k^{(d)})$, with ties resolved randomly. The ECC-calibrated sample $\widetilde{\boldsymbol{f}}^{(d)}$ for dimension d is obtained by rearranging the sample from step 1 according to $\boldsymbol{\pi}_d$; that is,

$$\widetilde{f}_k^{(d)} := \widehat{f}_{\pi_d(k)}^{(d)}, \quad k = 1, 2, \ldots, K, \quad d = 1, 2, \ldots, D.$$

A further non-parametric approach for multivariate post-processing examined here is the SSh (Clark *et al.*, 2004), which reconstructs dependence structures by reordering calibrated univariate samples to match the rank order of randomly selected historical observations of the same size. Here, we limit the sample size to that of the raw ensemble and apply the same sampling procedure for comparability with ECC; however, the method can generate ensembles of arbitrary size given a sufficiently long historical record. In this study, for both case studies, we consider the equidistant quantiles of the predictive distributions as input for both the ECC and the SSh methods.

3.3.1 | Graph neural networks

GNNs provide a flexible framework for learning from data defined on irregular relational structures and have been successfully applied across diverse domains, including modeling molecular structures and predicting molecular properties (Li *et al.*, 2025), as well as optimizing traffic networks and route planning (Jiang & Luo, 2022). Unlike traditional neural networks, such as convolutional neural networks, which assume inputs on regular grids or lattices (e.g., image pixels), GNNs operate directly on graphs, enabling them to capture complex interactions among nodes arranged arbitrarily and non-uniformly (Feik *et al.*, 2024).

A graph is formally represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V} = \{1, 2, \ldots, D\}$ corresponds to entities of interest, and the edges $\mathcal{E}$ encode relationships or interactions between them. Each node $d \in \mathcal{V}$ is associated with an M-dimensional feature vector $\mathbf{x}^{(d)} \in \mathbb{R}^M$, which may include relevant covariates, measurements, or summary statistics. By propagating and aggregating information along the graph edges, GNNs can capture both local and global dependencies, making them a versatile tool for tasks involving structured relational data.

3.4 | Verification metrics

Within the censored normal EMOS framework outlined in Section 3.1, model parameters are obtained by minimizing the mean of a proper scoring rule over the training dataset. The most widely adopted choice is the CRPS (Wilks, 2019, section 9.5.1), a standard verification metric extensively used in the atmospheric sciences.

For a predictive CDF F and an observation $y \in \mathbb{R}$, the CRPS is defined as

$$\begin{aligned}\text{CRPS}(F, y) &= \int_{-\infty}^{\infty} [F(x) - \mathbf{1}\{x \geq y\}]^2 \, dx \\ &= \mathbb{E}|X - y| - \frac{1}{2}\mathbb{E}|X - X'|,\end{aligned}$$

where X and X' are independent random variables with distribution F and finite first moments. Smaller CRPS values indicate superior predictive performance, as the score simultaneously captures calibration (statistical consistency between forecasts and observations) and sharpness (concentration of the predictive distribution). For the censored normal distribution, the CRPS admits a closed-form expression, allowing computationally efficient optimization and making it particularly suitable as a loss function in both EMOS and machine-learning frameworks.

Moreover, for an ensemble forecast $\{f_1, \ldots, f_K\}$ resulting in an empirical CDF $\hat{F}_K$ the previous equation reduces to the sample CRPS; that is,

$$\text{CRPS}(\hat{F}_K, y) = \frac{1}{K}\sum_{k=1}^{K}|f_k - y| - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K}|f_k - f_\ell|, \quad (1)$$

which is implemented, for example, in the `scoringRules` package in `R`.

For a confidence level $\alpha \in (0,1)$, the $[(1-\alpha)\times 100]\%$ central prediction interval, defined by the $\alpha/2$ and $1-\alpha/2$ quantiles, allows assessment of sharpness, measured by the average width, and calibration, quantified by the proportion of observations contained in the interval (coverage). For a K-member ensemble, setting $\alpha = 2/(K+1)$ aligns the nominal coverage with the empirical coverage $[(K-1)/(K+1)\times 100]\%$ of the raw ensemble (see e.g., Gneiting & Raftery, 2007).

The generalization of Equation (1) leads to the sample ES (Gneiting & Raftery, 2007). For an ensemble forecast $\mathbf{f}_k = (f_k^{(1)}, f_k^{(2)}, \ldots, f_k^{(D)})^{\mathrm{T}}$, $k = 1, \ldots, K$, and the corresponding observation vector $\mathbf{y} = (y^{(1)}, y^{(2)}, \ldots, y^{(D)})^{\mathrm{T}}$, the sample ES is given by

$$\mathrm{ES}(\hat{F}_K, \mathbf{y}) = \frac{1}{K}\sum_{k=1}^{K}\|\mathbf{f}_j - \mathbf{y}\| - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K}\|\mathbf{f}_k - \mathbf{f}_\ell\|,$$

where $\|\cdot\|$ denotes the Euclidean norm.

In addition to the ES, the multivariate performance of ensemble forecasts is also assessed using the VS of order p (Scheuerer & Hamill, 2015, VS_p;), which is particularly sensitive to errors in the correlation structure of the forecasts. The VS_p is defined as

$$\mathrm{VS}_p(F_K, \mathbf{y}) = \sum_{i=1}^{D}\sum_{j=1}^{D}\omega_{ij}\left(|y^{(i)} - y^{(j)}|^p - \frac{1}{K}\sum_{k=1}^{K}|f_k^{(i)} - f_k^{(j)}|^p\right)^2,$$

where $\omega_{ij} \geq 0$ are weights that quantify the relative importance of each coordinate pair (i, j).

Common choices for the order p are 0.5 and 1. In our study, we use $p = 0.5$ and denote the score simply as VS.

Graphical tools for assessing the reliability of multivariate forecasts, such as multivariate rank histograms, are also well established. These histograms are based on pre-ranks, which condense a multivariate quantity into a single value. Since interpreting multivariate histograms can be challenging, it is advisable to employ multiple pre-ranking methods concurrently to obtain a comprehensive view of forecast miscalibration (Allen *et al.*, 2024). Following this strategy, we consider average-, band-depth-, ES-, and dependence-based histograms in our analysis. The average rank is computed as the mean of the ranks of observations within the ensembles. Its histogram offers a diagnostic assessment of the forecast distribution: underdispersion is reflected by a ∪-shaped histogram, overdispersion by a ∩-shaped histogram, and bias typically results in a triangular pattern, whereas the band depth rank of each observation reflects its centrality relative to the ensemble of forecasts Thorarinsdottir *et al.* (2016). Whereas the ES and dependence histograms build on the more recent concept of proper score-based pre-ranks introduced by Knüppel *et al.* (2022).

In addition, the reliability index (RI) provides a quantitative measure of the uniformity of a rank histogram, thereby indicating the degree to which the ensemble is properly calibrated. For a histogram with $K+1$ bins and predicted frequencies $\hat{p}_k$ in each bin, the RI is given by

$$\mathrm{RI} = \sum_{k=1}^{K+1}\left|\hat{p}_k - \frac{1}{K+1}\right|,$$

where smaller values of RI indicate a more uniform histogram.

The statistical significance of differences in verification scores is assessed using the Diebold–Mariano (DM) test (Diebold & Mariano, 1995), which accounts for temporal dependencies in forecast errors. Given a scoring rule and two competing forecasts, F and F_{ref}, the DM test statistic is defined as

$$t = \frac{\bar{d}}{\hat{\sigma}_{\bar{d}}/\sqrt{n}},$$

where n denotes the number of forecast cases in the test set, $\bar{d}$ is the mean difference in scores over the test set between forecasts F and F_{ref}, and $\hat{\sigma}_{\bar{d}}$ is a consistent estimator of the asymptotic standard deviation of the individual score differences. Under standard regularity conditions, the DM statistic asymptotically follows a standard normal distribution under the null hypothesis of equal predictive accuracy. Negative values of the DM statistic indicate better predictive performance of F, whereas positive values favor F_{ref}. To account for multiple testing, the Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) is applied.

In the case studies presented in Section 5, improvements of forecasts relative to a reference forecast F_{ref} are quantified using skill scores derived from a verification metric S (Gneiting & Raftery, 2007). The skill score is defined as

$$\text{Skill score} = 1 - \frac{\bar{S}_F}{\bar{S}_{F_{\mathrm{ref}}}},$$

where $\bar{S}_F$ and $\bar{S}_{F_{\mathrm{ref}}}$ denote the mean values of the score S over all forecast cases in the verification period for the forecast F and the reference F_{ref}, respectively. When S corresponds to the CRPS, this skill score is commonly referred to as the continuous ranked probability skill score (CRPSS). In addition to the CRPSS, we evaluate forecasts using the energy skill score (ESS) and the variogram skill score (VSS), with higher values indicating better performance.

4 | IMPLEMENTATION DETAILS

4.1 | GNN methods

For the GNN model considered here, the spatial structure of the observational network is represented as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V} = \{1, 2, \ldots, D\}$ corresponds to the observation stations, and the edges $\mathcal{E}$ capture spatial relationships. An edge is established between two nodes v and v' if the Haversine distance between the corresponding stations is below a threshold r, which is selected to minimize the CRPS. The graph yielding the best performance for both datasets is shown in Figure 1. The threshold is set to $r = 50$ km for solar irradiance and $r = 100$ km for visibility.

In both case studies, we use the GraphSAGE architecture (Hamilton *et al.*, 2017) with the mean aggregator, differing only in minor hyperparameter settings. For solar irradiance, the network consists of a single hidden `SAGEConv` layer (responsible for aggregating neighborhood information) mapping the input features to a hidden representation, followed by batch normalization, rectified linear unit activation, and dropout. For visibility, the network uses two hidden `SAGEConv` layers with the same set-up for each layer. In both cases, a final `SAGEConv` layer outputs the forecast-specific representation without an activation function. In principle, a non-negativity constraint could be enforced through activation functions, which would be relevant for both of the meteorological variables considered. However, to ensure architectural consistency with the MLP of Baran *et al.* (2025), the non-negativity requirement was implemented within the cost functions instead.

For both case studies, a single model was initialized to process all lead times jointly. For all GNN models, we applied a rolling training window, which means that the n days preceding the forecast date were used to create the training data. For solar irradiance, $n = 30$, whereas $n = 530$ for visibility. Hence, the training set consists of a sequence of daily graphs, where each graph is defined by its node features and edge structure. However, in the present work, only the node features vary over time, whereas the edge connections remain fixed. For solar irradiance forecasting, the first 24 lead times are processed in a single pass, which results in $24n$ training graphs. After the prediction for those lead times of the forecast date is completed, another cycle is performed using the next set of $24n$ graphs, to provide predictions for the next 24 hr. As a result, the network generates node-wise solar irradiance forecasts for each lead time, providing 24-hr-ahead predictions for all stations in the graph. For visibility forecasting, where 20 lead times are available, the training set contains $20n$ graphs per forecast date. These graphs are then divided into training and validation subsets (70/30 split), ensuring that both subsets contain complete graphs sampled from the same rolling window. The test dataset for the forecast date consists of a set of unlabeled graphs corresponding to a single day, where each graph represents one lead time processed by the model in a single pass, on which the trained model produces the predictions for the respective lead times. As mentioned earlier, here we utilize the mean aggregator, which aggregates features from a node's local neighborhood by taking the element-wise mean of the neighbor feature vectors, as illustrated in Figure 2. During training, the network weights are updated based on the node's own features and the aggregated neighborhood features (Hamilton *et al.*, 2017). Training was conducted for up to 500 epochs, employing early stopping based on validation loss with a patience of 15 epochs for solar irradiance and 10 epochs for visibility. The hyperparameters of the GNN models are summarized in Table 1.

For solar irradiance, each observation station is characterized by features including the ensemble mean, the relative frequency of zero irradiance values, and the ensemble variance for the given lead time and day. Additionally, the station's geographical coordinates (latitude, longitude), elevation, and the forecast lead time are included to distinguish different forecast hours within the day. This feature set is fully consistent with that used in Baran *et al.* (2025).

For visibility, each observation station is represented by a feature set that includes the control forecast, the mean and standard deviation of the 50 exchangeable ensemble members, and the proportions of forecasts predicting visibility below 5000 m, between 5000 and 30,000 m, and 30,000–70,000 m. Additionally, visibility point forecasts from the Copernicus Atmosphere Monitoring Service are incorporated, along with annual base functions to capture seasonal and temporal effects, defined as

$$\beta_1(d) := \sin\left(\frac{2\pi d}{365}\right) \quad \text{and} \quad \beta_2(d) := \cos\left(\frac{2\pi d}{365}\right),$$

where d denotes the day of the year. Furthermore, the station's latitude, longitude, and elevation are also taken into account. As with the solar irradiance data, this feature set largely follows Lakatos and Baran (2024); however, for the GNN, the geographical coordinates are explicitly included to enable graph construction and capture location-dependent information. Lead time is also included to account for temporal variations.

Hyperparameters were tuned in an iterative manner on a single NVIDIA GeForce RTX 3060. Owing to computational constraints, we first determined the optimal training window length while keeping other settings fixed, then we optimized the number of layers and proceeded similarly for the remaining hyperparameters. To assess uncertainty,

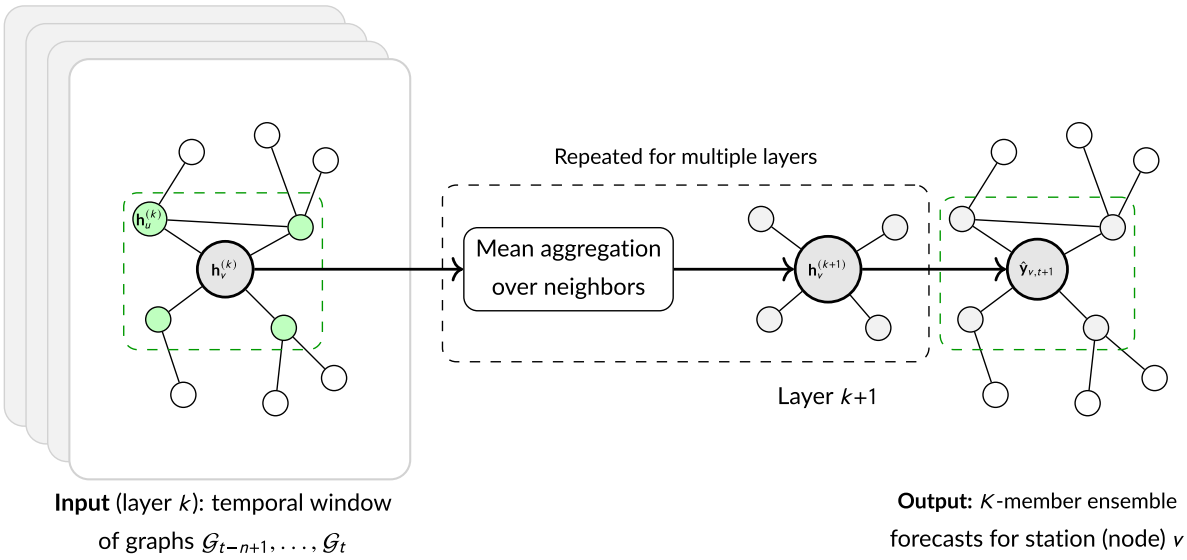

FIGURE 2 Simplified illustration of the GraphSAGE model with mean aggregation. The input is a temporal window of past n days (graphs) $\mathcal{G}_{t-n+1}, \ldots, \mathcal{G}_t$, and the model produces node-wise K-member ensemble forecasts. For illustration purposes, the figure considers a single lead time (one representative node is shown). Here, v denotes the target node, u denotes a neighboring node of v, $\mathbf{h}_v^{(k)}$ and $\mathbf{h}_u^{(k)}$ represent the feature vectors of the target node and its neighbors at layer k derived from the temporal window ending at day t, and $\mathbf{h}_v^{(k+1)}$ denotes the updated representation of the target node at layer $k+1$. Finally, $\hat{\mathbf{y}}_{v,t+1}$ denotes the K-member ensemble forecast produced for node v for the forecast day $t+1$. (The exact temporal shift depends on the selected lead time and is not restricted to $t+1$). [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Hyperparameter settings of the graph neural network models for solar irradiance and visibility forecasting.

Parameter	Solar irradiance	Visibility
Number of nodes/ stations D	18	30
Edge definition	Distance < 50 km	Distance < 100 km
Aggregator	Mean aggregator	Mean aggregator
Number of hidden `SAGEConv` layers	1	2
Number of hidden units	1,024	64
Output dimension K	8	51
Dropout	0.2	0.2
Batch size	64	64
Learning rate	0.03	0.03
Validation set size	0.3	0.3
Number of epochs	500	500

each model was trained and evaluated 10 times during testing, with performance measured by the average verification scores across these runs. The models generally exhibited stable behavior, with only minor performance differences observed between iterations.

The hyperparameters listed in Table 1 were first found by minimizing the CRPS. Subsequently, multivariate loss functions were introduced to enable the GNN to more effectively capture dependencies between stations. Specifically, apart from the variant trained solely by minimizing the ES, we employed a weighted combination of the ES and the VS:

$$\mathcal{L} = w_1 \cdot \mathrm{ES} + w_2 \cdot \mathrm{VS},$$

where w_1 and w_2 are weighting coefficients. In our experiments, we set $w_1 \in [0, 1]$ and $w_2 = 1 - w_1$, for simplicity and interpretability, although in principle other values (e.g., $w_1 > 1$) could be used. Owing to the large difference in scale between the two terms (the VS was found empirically to be several orders of magnitude larger than the ES) the VS component was normalized by an appropriate factor. Specifically, this factor was computed as the ratio of the mean ES to the mean VS over the raw ensemble, which was consistent with the ratios observed in the training batches. This normalization allowed us to select a weighting within the $[0, 1]$ interval that balances the contributions of both terms, ensuring that neither dominates the loss and that optimization proceeds stably. In the case of solar irradiance forecasts, excessively high ES weights reduced the proportion of better VS cases, and high VS weights degraded ES performance and CRPS. The 0.9–0.1 ES–VS weighting achieved the best balance, simultaneously maximizing the proportions of significantly better

ES and VS cases and maintaining reasonable mean CRPS. Notably, adding the VS component did not compromise ES performance but substantially improved VS and also ES values compared with GNN models trained with only ES or CRPS loss. During dualGNN testing with identical hyperparameters, the 50 km threshold model outperformed the unconnected reference in ES for 20 of 26 lead times and in VS for five lead times. Supporting Information Figure S1 illustrates the advantage of selecting an optimal threshold over graphs with too few or too many edges. For visibility, the dualGNN minimizes ES with a weight of 0.3 and VS with a weight of 0.7. As with the radiation data, increasing the weight of ES improves ES performance, whereas placing greater weight on VS enhances VS performance. The chosen weighting scheme not only provides a balanced compromise but also yields the lowest mean CRPS among all dualGNN variants tested with different weightings. For the visibility forecast, we performed the same analysis as for the solar irradiance data to evaluate how connectivity influences the spatial behavior of the models. For both ES and VS, the 100 km threshold model performed best, outperforming the "edgeless" dualGNN in ES for five of 20 lead times and in VS for 15 lead times, whereas it was never outperformed by any other threshold-based model. A comparison of model performances is shown in Supporting Information Figure S2.

The GNN models were implemented in Python using the `PyTorch Geometric` library (Fey & Lenssen, 2019).

4.2 | Comparison of the GNN and benchmark methods

For the post-processing of solar irradiance forecasts, we employed the MLP model previously used by Baran *et al.* (2025) on the same dataset, adopting the same hyperparameters as in their study. Specifically, a 25-day rolling training period was used, with a batch size of 1200 and a learning rate of 0.01. The input features were identical to those summarized earlier, and the network architecture consisted of two hidden layers with 255 neurons each. Early stopping was implemented with a patience of five epochs, and the test set comprised 20% of the data.

For the EMOS model, we applied the semi-local approach of Lerch and Baran (2017), briefly described in Section 3.1. In this case, the 18 stations were grouped into three clusters, each containing at least two stations. The rolling training period was set to 80 days based on an extensive hyperparameter search. Although GNNs typically require larger datasets for effective parameter estimation, a 30-day training period yielded the best performance in this case, likely due to maintaining relevance to the current atmospheric conditions. Though one of the main advantages of EMOS lies in its computational efficiency, the number of parameters to be estimated increases to 69,368 and 2,134,024 for the MLP and the GNN models respectively. Although these models are not directly comparable to EMOS, since EMOS treats each lead time separately, it is clear that, in terms of parameter count, the GNN faces the greatest disadvantage. Both the MLP and the dualGNN generate eight-member ensemble forecasts to ensure comparability with the raw WRF forecasts. For the same purpose, we also consider eight equidistant quantiles extracted from the censored normal EMOS predictive distributions, with these samples hereafter referred to simply as EMOS. For a more comprehensive evaluation, we also consider GNN variants optimizing only ES or CRPS (denoted GNN-ES and GNN-CRPS) alongside the dualGNN. In this study, the same set of hyperparameters was used for all three GNN variants (GNN-ES, GNN-CRPS, and dualGNN). The hyperparameters were selected by CRPS minimization, and the resulting architectures were subsequently tested using ES and the dual loss, allowing assessment of the contribution of each loss function to overall model performance.

Finally, the reference forecasts used as the basis for the visibility dualGNN were generated by the POLR classifier described in Section 3.2. To ensure consistency with the dualGNN outputs, a sample of size 51 was drawn from the predictive probability mass functions (PMFs) produced by these classifiers. These models were trained over a 350-day period and used the same set of features as in Lakatos and Baran (2024). As the GNN models consider the classes independent, the POLR provides a strong baseline, as it is specifically designed for ordered data such as the visibility observations considered here. Similar to the solar irradiance predictions, the GNN models generate forecasts that are both calibrated and spatially consistent, using the same discrete observational categories as POLR. Unlike POLR, which requires a maximum to be imposed on values derived from the PMFs, the GNN models can produce values of any magnitude, unconstrained by the historical observation range. The GNN models utilize 530 days of training data. Consistent with the approach in Section 5.1, we evaluate three variants of the GNN models: the dualGNN, which optimizes a composite loss function, and the GNN-CRPS and GNN-ES models, which minimize the CRPS and the ES respectively. Following a similar approach to the EMOS models for post-processing solar irradiance, the POLR forecast PMFs are represented by drawing 51 equidistant quantiles, enabling direct comparison with both the raw ensemble forecasts and the GNN-generated samples. As in Lakatos and Baran (2024), to enable a fair comparison between the raw and all post-processed forecasts, POLR samples were randomized to generate continuous values. As mentioned

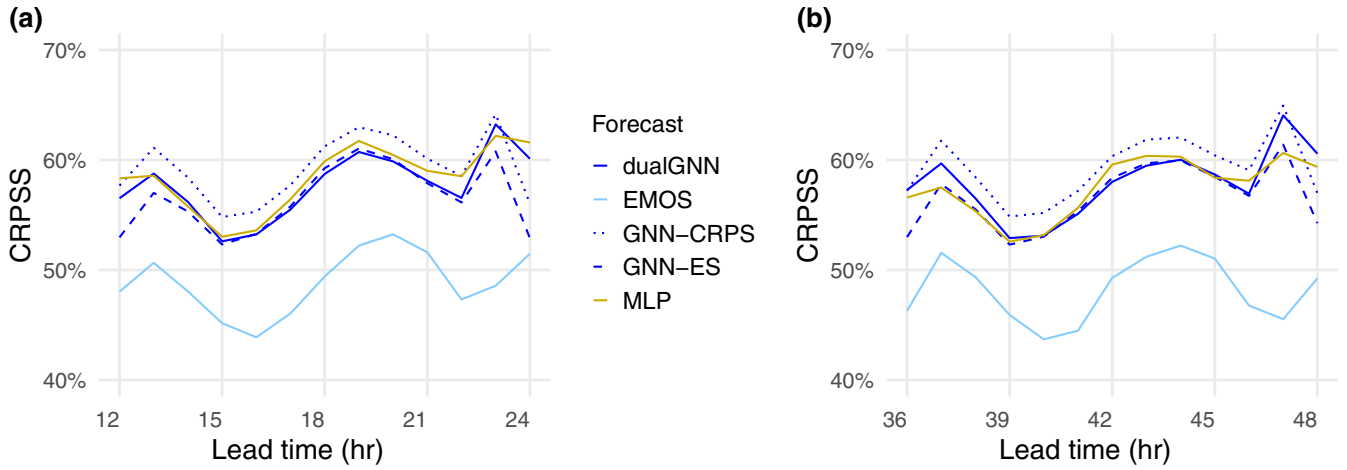

FIGURE 3 Continuous ranked probability skill score (CRPSS) of ensemble model output statistics (EMOS)- and graph neural network (GNN)-based post-processed solar irradiance forecasts relative to the raw ensemble as functions of the lead times (a) 12–24 hr and (b) 36–48 hr. CRPS: continuous ranked probability score; ES: energy score; MLP: multilayer perceptron. [Colour figure can be viewed at wileyonlinelibrary.com]

in Section 3.2, in terms of the number of parameters to be estimated, the statistical POLR model clearly has an advantage, whereas for the GNN models the number of parameters to be estimated increases to 25,203.

5 | RESULTS

We now present the results of post-processing applied to forecasts of solar irradiance and visibility. Though our main focus is on improving the representation of spatial dependencies between locations, we also provide a brief assessment of the marginal forecast accuracy for each variable, which enables evaluation of both individual performance and the effectiveness in capturing spatial correlations across observation sites.

5.1 | Univariate performance of solar irradiance forecasts

Figure 3 displays the CRPSS values relative to the raw ensemble forecasts. As shown in Figure 3a, all post-processed forecasts substantially improve upon the raw ensemble for lead times 12–24 hr, which are particularly relevant for solar irradiance applications. During this period, the EMOS model provides the least improvement, whereas the dualGNN, GNN-ES, and MLP models perform nearly identically, and the GNN-CRPS achieves the highest improvement upon to the raw ensemble. The overall ranking of forecasts based on Figure 3b for the next day's daylight hours is essentially the same. This ranking becomes even clearer when examining Table 2,

TABLE 2 Overall mean continuous ranked probability skill score (CRPSS) of post-processed solar irradiance forecasts with respect to the raw ensemble for lead times 12–24 hr and 36–48 hr.

Method	CRPSS (%)
EMOS	48.54
MLP	57.95
dualGNN	57.79
GNN-ES	56.56
GNN-CRPS	**59.25**

Abbreviations: CRPS: continuous ranked probability score; GNN: graph neural network; EMOS: ensemble model output statistics; ES: energy score; MLP: multilayer perceptron.
Note: The value in bold indicates the greatest improvement with respect to the reference.

which reports the overall mean CRPSS of post-processed solar irradiance forecasts with respect to the raw ensemble for lead times 12–24 hr and 36–48 hr. Based on this metric, the GNN-CRPS model yields the largest improvement, followed by the MLP, and closely by the dualGNN. Supporting Information Figure S3, as an extension of the previous results, also shows the performance of the models during the dark hours.

Table 3 displays the coverage and average width of the nominal 77.78% central prediction intervals for post-processed and raw solar irradiance forecasts across lead times. Here, the coverage of the entire ensemble range is considered. All post-processed forecasts exhibit better calibration than the raw ensemble forecasts. Considering the mean absolute deviation from the nominal coverage across all 48 lead times, the dualGNN is closest to the nominal level, followed closely by the MLP (not shown). When focusing on forecasts for the 12–24 hr and 36–48 hr lead

TABLE 3 Coverage and average width of the 77.78% central prediction intervals for post-processed and raw solar irradiance forecasts for the 12–24 hr and 36–48 hr lead times.

	Ensemble	EMOS	dualGNN	GNN-CRPS	GNN-ES	MLP
Average width, 12–24 hr	62.5762	173.7423	189.1103	184.0687	200.6200	219.9254
Average width, 36–48 hr	63.6587	179.9249	191.2580	186.9103	202.7045	177.6645
Coverage, 12–24 hr (%)	27.58	71.11	79.47	76.42	79.76	80.11
Coverage, 36–48 hr (%)	27.24	70.98	79.51	76.41	79.71	76.34

Abbreviations: CRPS: continuous ranked probability score; EMOS: ensemble model output statistics; ES: energy score; GNN: graph neural network; MLP: multilayer perceptron.

times, the GNN-CRPS is closest to the nominal level, followed by dualGNN for 12–24 hr and MLP for 36–48 hr. As generally expected, improved calibration is associated with wider prediction intervals, with the MLP producing the widest intervals for the first 24 hr, followed by the GNN models and EMOS. In the second 24 hr the order of machine-learning models reverses, and the GNN models produce slightly wider intervals than the MLP. Overall, the GNN intervals remain relatively consistent across both periods, whereas the MLP intervals narrow slightly in the second 24 hr, approaching those of EMOS. Supporting Information Figure S4 provides a graphical illustration of the coverage and average width values, also taking the evening hours into account.

5.2 | Univariate performance of visibility forecasts

Figure 4 shows the CRPSS of post-processed forecasts relative to the POLR predictions as functions of the lead time. As in the previous case study, the results summarized in Table 4 indicate that all post-processed models outperform the raw ensemble forecasts. However, based on Supporting Information: Figure S5, this advantage slightly decreases with increasing lead time for all models. Differences in model performance are more clearly visible based on Table 4, that quantifies the improvement with respect to the reference POLR model. The GNN-CRPS model provides the largest improvement relative to POLR, followed by the dualGNN, while the GNN-ES model also shows positive gains over POLR.

Finally, Table 5 presents the coverage and average widths of the 90% central prediction intervals for raw and post-processed visibility forecasts as functions of lead time. The dualGNN model exhibits a mean absolute deviation of 0.89% from the nominal 90% coverage, compared to 2.3% for POLR and 50.81% for the raw ensemble. Moreover, based on Supporting Information Figure S6, temporal dependencies and the diurnal cycle are less pronounced in the coverage values of all post-processed models than

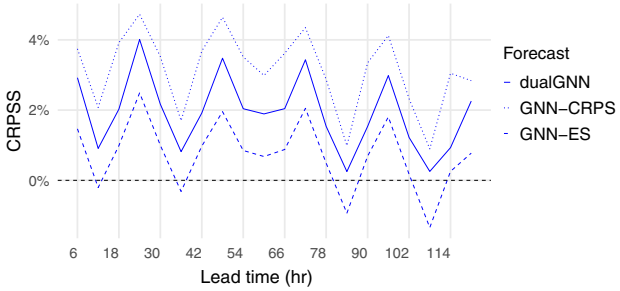

FIGURE 4 Continuous ranked probability skill score (CRPSS) of post-processed forecasts relative to the proportional odds logistic regression predictions as functions of the lead time. CRPS: continuous ranked probability score; ES: energy score; GNN: graph neural network. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Overall mean continuous ranked probability skill score (CRPSS) of raw and post-processed visibility forecasts with respect to the proportional odds logistic regression forecasts.

Model	CRPSS (%)
Ensemble	−54.15
dualGNN	1.93
GNN-CRPS	**3.15**
GNN-ES	0.74

Abbreviations: CRPS: continuous ranked probability score; ES: energy score; GNN: graph neural network.
Note: The value in bold indicates the greatest improvement with respect to the reference.

in the raw ensemble. The raw ensemble produces the narrowest prediction intervals, whereas the GNN models have narrower average widths than POLR and are less sensitive to the diurnal cycle.

5.3 | Multivariate performance of solar irradiance forecasts

In this section, the performance of the proposed GNN models is assessed in comparison with a range of empirical copula-based approaches. Specifically, when the EMOS

TABLE 5 Mean coverage and average widths of 90% central prediction intervals of raw and post-processed visibility forecasts.

	POLR	Ensemble	MLP	dualGNN	GNN-CRPS	GNN-ES
Average width	45570.52	22786.83	54790.92	43785.37	42955.40	42552.63
Coverage (%)	87.70	39.19	91.32	89.11	89.06	87.87

Abbreviations: CRPS: continuous ranked probability score; ES: energy score; GNN: graph neural network; MLP: multilayer perceptron; POLR: proportional odds logistic regression.

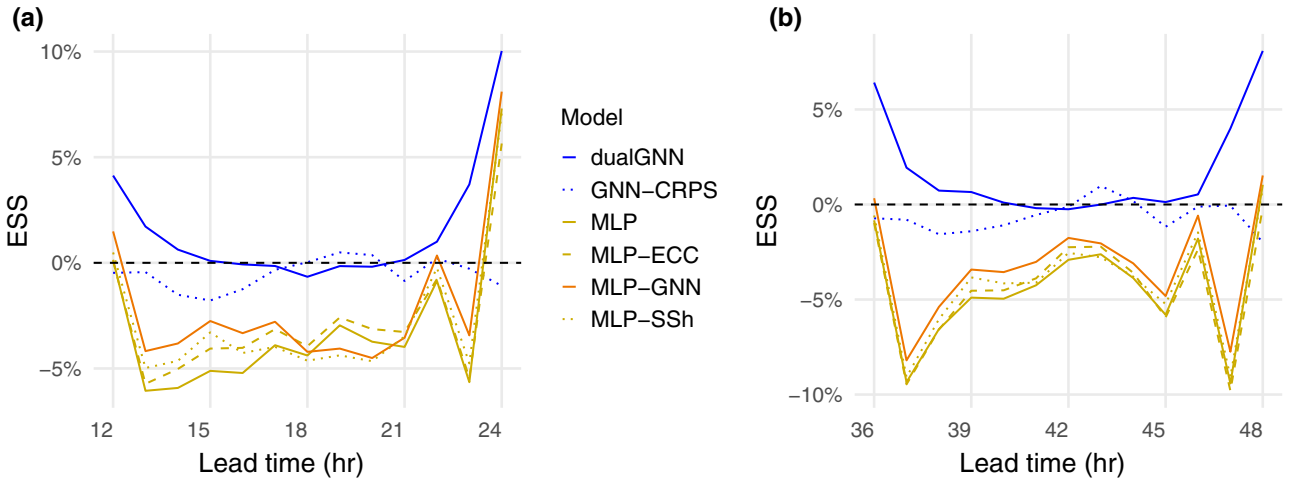

FIGURE 5 Energy skill score (ESS) of post-processed solar irradiance forecasts relative to the GNN-ES model as functions of the lead times (a) 12–24 hr and (b) 36–48 hr. CRPS: continuous ranked probability score; ECC: ensemble copula coupling; GNN: graph neural network; MLP: multilayer perceptron; SSh: Schaake shuffle. [Colour figure can be viewed at wileyonlinelibrary.com]

and MLP samples are reordered to match the rank structure of the raw WRF forecasts, the resulting methods are denoted EMOS-ECC and MLP-ECC, respectively. Alternatively, when the dependence template is derived from past observations within the training period, the approaches are referred to as EMOS-SSh and MLP-SSh. As shown by Lakatos *et al.* (2023) restricting the set of possible pool dates to the training period does not deteriorate forecast performance. Furthermore, for the current dataset, we include in the analysis a hybrid model that combines the MLP with the dualGNN by mapping the MLP-generated samples onto the rank order structure of the dualGNN forecasts for the same verification day and forecast horizon. Consequently, this hybrid model requires running the MLP and dualGNN in parallel. This reference model, hereafter referred to as MLP-GNN, enables us to evaluate whether the dualGNN rank structure provides additional information beyond that contained in the raw ensemble forecasts and historical observations. To provide a baseline, we also consider the EMOS and MLP samples under the assumption of independence, meaning that the calibrated samples are not rearranged according to any dependence template.

Figure 5 displays the mean ESS of post-processed solar irradiance forecasts relative to GNN-ES as a function of the 12–24 h (a) and 36-48 h (b) lead times. The raw and EMOS-based forecasts are omitted to better visualize the performance of the higher-ranking models. As shown in the figure, the ranking of the models remains consistent between the first and second day. Based on the first row of Table 6, which quantifies the overall ESS for lead times of 12-24 h and 36-48 h, the models are ranked as follows: dualGNN, GNN-CRPS, MLP-GNN, MLP-ECC, MLP-SSh, MLP, EMOS-SSh, EMOS-ECC, EMOS, and finally the raw ensemble. This ranking demonstrates that integrating VS into the training objective enhances ES performance. Furthermore, Figure 5b shows that dualGNN is the only model achieving a positive gain relative to GNN-ES for the 36-48 h forecasts, whereas for the first-day forecast horizon (panel a) the competition among the models is more balanced. It can also be observed that for both days, during periods of highest solar irradiance (lead times 18-20 h and 42-44 h), GNN-CRPS outperforms dualGNN. Overall, for both days, the effect of including the VS in the loss function is most pronounced at the beginning and end of the day, i.e., during periods of low solar irradiance. This suggests that the benefits of VS are particularly evident under low-irradiance conditions, where the spatial dependence captured by VS contributes most to improving forecast skill.

TABLE 6 Overall mean energy skill score (ESS) and variogram skill score (VSS) of the raw and post-processed solar irradiance forecasts with respect to the GNN-ES for lead times 12–24 hr and 36–48 hr.

	Ensemble	EMOS	EMOS-ECC	EMOS-SSh	MLP	MLP-SSh	MLP-ECC	MLP-GNN	GNN-CRPS	dualGNN
ESS (%)	−112.75	−20.32	−18.23	−17.31	−3.10	−2.75	−2.72	−2.5	−0.53	**1.56**
VSS (%)	−191.88	−58.45	−37.45	−28.62	−2.42	−0.42	−1.60	0.54	−2.26	**4.15**

Abbreviations: CRPS: continuous ranked probability score; ECC: ensemble copula coupling; EMOS: ensemble model output statistic; GNN: graph neural network; MLP: multilayer perceptron; SSh: Schaake shuffle.
Note: The value in bold indicates the greatest improvement with respect to the reference.

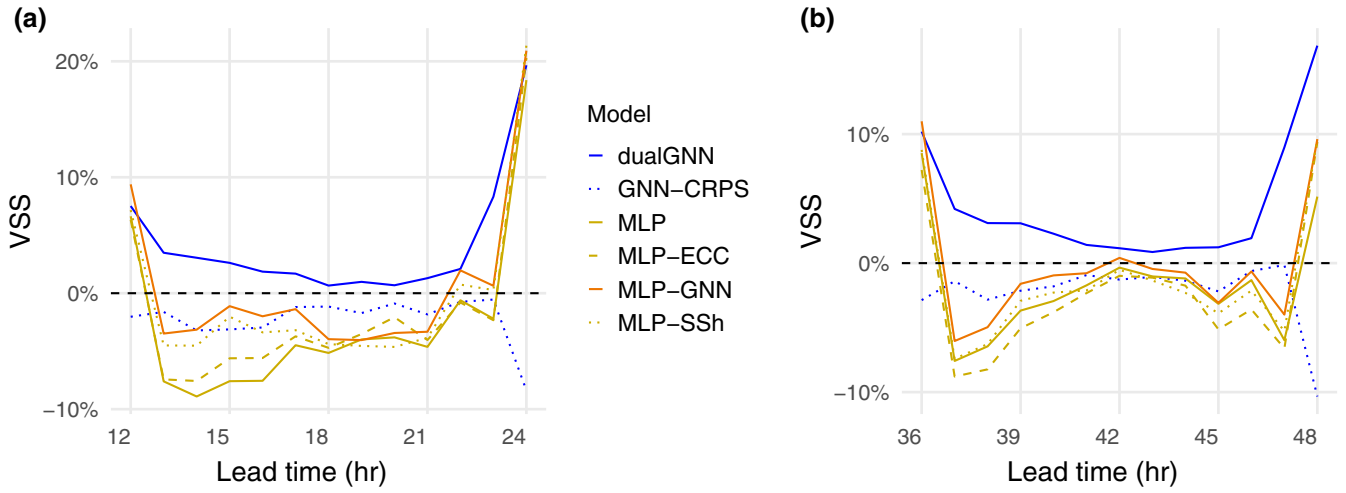

FIGURE 6 Variogram skill score (VSS) of post-processed solar irradiance forecasts relative to the GNN-ES as functions of the lead times (a) 12–24 hr and (b) 36–48 hr. CRPS: continuous ranked probability score; ECC: ensemble copula coupling; GNN: graph neural network; MLP: multilayer perceptron; SSh: Schaake shuffle. [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 6, which shows the mean VSS of post-processed solar irradiance forecasts relative to GNN-ES as a function of the 12–24 hr (Figure 6a) and 36–48 hr (Figure 6b) lead times, presents a very similar pattern. As in Figure 5, the raw and EMOS-based forecasts are omitted from the figure to better highlight the performance differences among the higher performing multivariate models. When forecasts are ranked according to the overall VSS during lead times of 12–24 hr and 36–48 hr, as quantified in the second row of Table 6, the models are ordered as follows: dualGNN, MLP-GNN, MLP-SSh, MLP-ECC, GNN-CRPS, MLP, EMOS-SSh, EMOS-ECC, EMOS, and finally the raw ensemble. Similar to Figure 5, the model ranking is fairly consistent across the two consecutive days. Relative to GNN-ES, based on the overall mean VSS values of Table 6 during lead times corresponding to 12–24 hr and 36–48 hr, the EMOS model exhibits an average deficit of 58.45%, followed by EMOS-ECC with 37.45% and EMOS-SSh with 28.62%. In contrast, MLP and GNN-CRPS models show much smaller deficits of 2.42% and 2.26% respectively, whereas the MLP-ECC has a deficit of 1.6%. The MLP-SSh has a small deficit of 0.42%, whereas the model reordered according to the GNN rank structure shows a 0.54% advantage. Finally, the dualGNN achieves an average improvement of 4.15% over GNN-ES. Based on these results, as also observed in Figure 5, the model performances are closest to the reference model during daytime periods with higher solar irradiance, meaning that their advantage or deficit relative to the reference is smallest in these intervals. Similar to the results shown in Figure 5, the advantage of dualGNN is most pronounced in second-day forecasts, particularly during the early evening hours. Nevertheless, dualGNN consistently outperforms the other models throughout daytime periods as well. As an extension of the previous results, Supporting Information Figures S7 and S8 present the mean ES and VS values, together with the corresponding skill scores, taking the nighttime hours into account.

Figure 7 presents boxplots of DM test statistics evaluating the significance of differences in ES relative to the strongest GNN-independent reference, MLP-ECC, and the GNN-ES method as functions of lead time. The gray shaded area indicates the acceptance region of the two-tailed DM test for equal predictive performance at a 5% significance level, and the Benjamini–Hochberg procedure was applied to control the false discovery rate. Negative DM values correspond to better predictive skill compared with the reference.

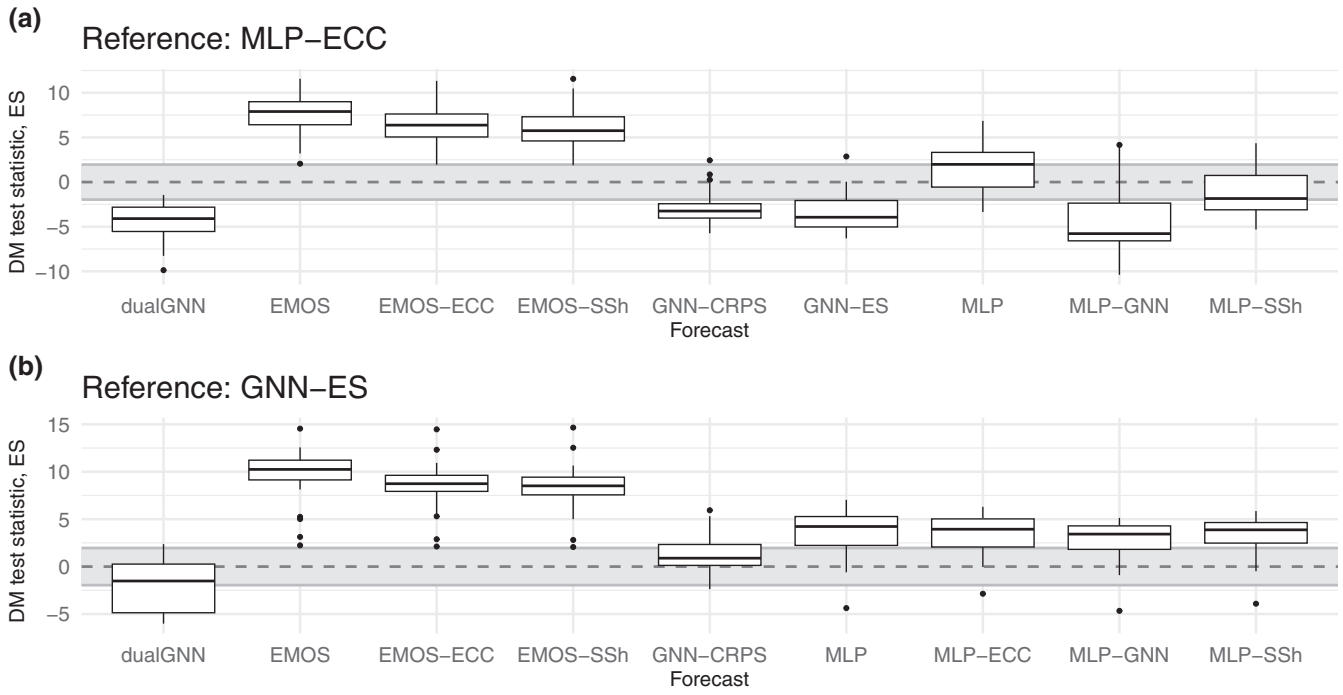

FIGURE 7 Boxplots of Diebold–Mariano (DM) test statistics investigating the significance of the difference in energy score (ES) (a) from the reference MLP-ECC and (b) GNN-ES methods as functions of the lead times 12–24 hr and 36–48 hr. Gray region indicates the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance, and negative values indicate better predictive performance compared with the reference. CRPS: continuous ranked probability score; ECC: ensemble copula coupling; EMOS: ensemble model output statistic; GNN: graph neural network; MLP: multilayer perceptron; SSh: Schaake shuffle.

As shown in Figure 7a, dualGNN, MLP-GNN, GNN-ES, and GNN-CRPS exhibit the greatest advantages, significantly outperforming the GNN-independent reference. Specifically, dualGNN performs better than the reference in 88.46% of cases (23 lead times out of 26) and is the only model that is never outperformed by the MLP-ECC reference model, whereas all other methods fall short compared with MLP-ECC for at least two lead times. For the other three models, this proportion is 76.92%. Following these are MLP-SSh and MLP, which achieve improvements in 46.15% and 15.38% of cases respectively. All EMOS variants, however, are outperformed by MLP-ECC in 96.15–100% of cases.

In Figure 7b, a somewhat different picture emerges. In this case, dualGNN outperforms the ES-trained GNN in 46.15% of cases, whereas in 50% of cases the null hypothesis of equal predictive performance cannot be rejected. For all MLP variants, including MLP-GNN, the improvement is limited to only 3.84%; among these, MLP-GNN most frequently achieves performance that is statistically comparable to the baseline. GNN-CRPS follows in the ranking: although it never surpasses the ES-trained model, its performance coincides with it in 79.92% of cases. By contrast, all EMOS variants are consistently outperformed by the ES-trained GNN for all lead times. Based on the plot analyzing VS deviations, similar conclusions can be drawn as in the case of Figure 7, which is presented as Supporting Information Figure S10 and briefly discussed there.

Figure 8 presents the multivariate rank histograms of the post-processed forecasts, with the associated RIs indicated on the plots, where the lowest RIs are highlighted in bold. Though the RI provides a convenient summary of deviations from uniformity, interpreting multivariate rank histograms requires caution, as their meaning depends on the choice of pre-ranking method. In general, the two-step post-processing models and the GNN-based approaches produce the most uniform average- and band-depth rank histograms, with dualGNN exhibiting the most even distributions. Based on the average ranks, the independent EMOS model achieves the highest RI, and the MLP-ECC model shows the highest value according to the band-depth histogram. All models show a left-skew in the ES rank histograms, with the effect being least pronounced for GNN-ES. Based on the VS-based dependence rank histograms, the histogram corresponding to the MLP reordered according to the GNN rank structure is the most uniform, whereas deviations in the correlation structure would manifest as right- or left-skewed histograms.

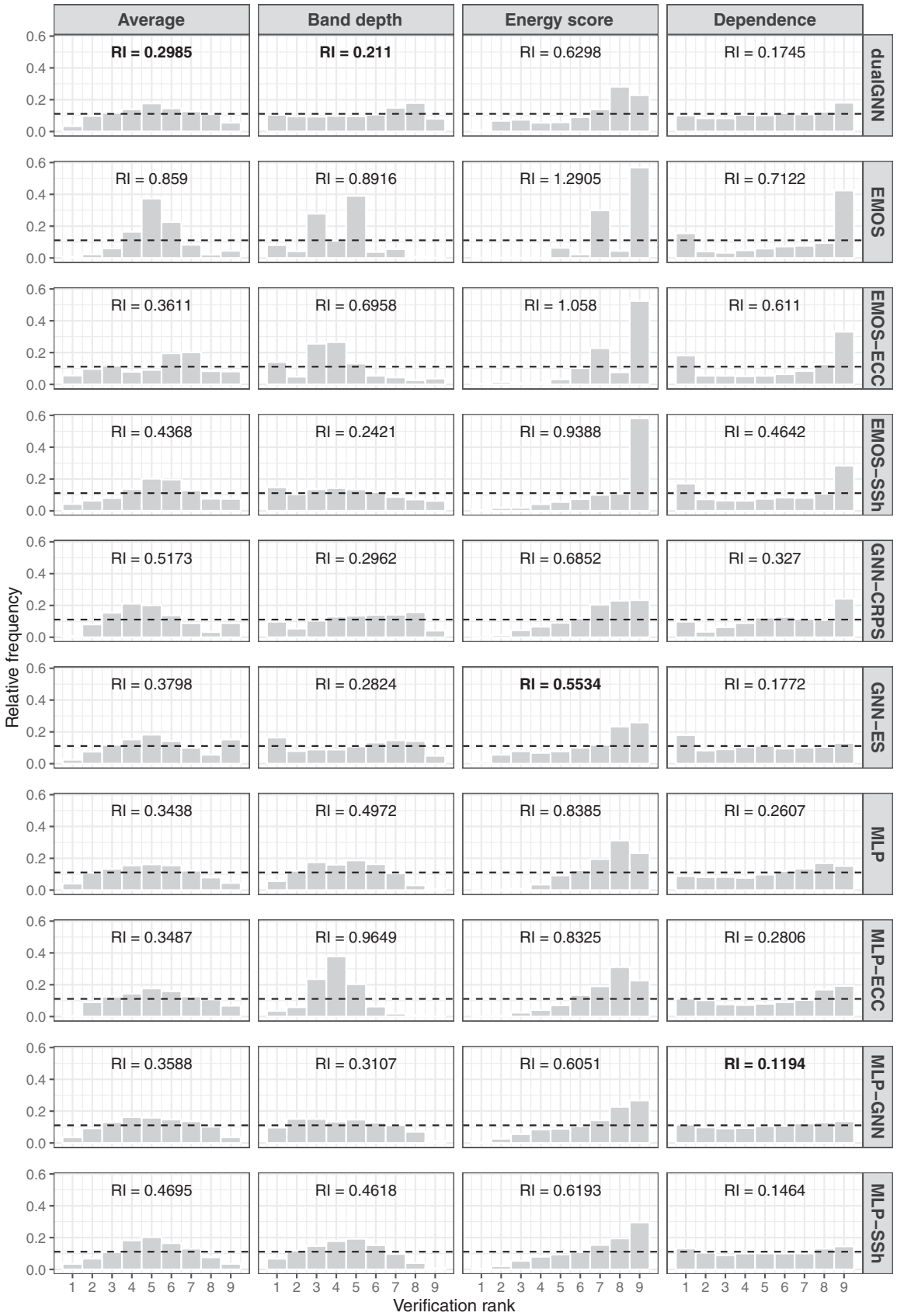

FIGURE 8 Rank histograms of post-processed solar irradiance forecasts together with the corresponding reliability indices (RI) for lead times 12–24 hr and 36–48 hr. CRPS: continuous ranked probability score; ECC: ensemble copula coupling; EMOS: ensemble model output statistic; ES: energy score; GNN: graph neural network; MLP: multilayer perceptron; SSh: Schaake shuffle.

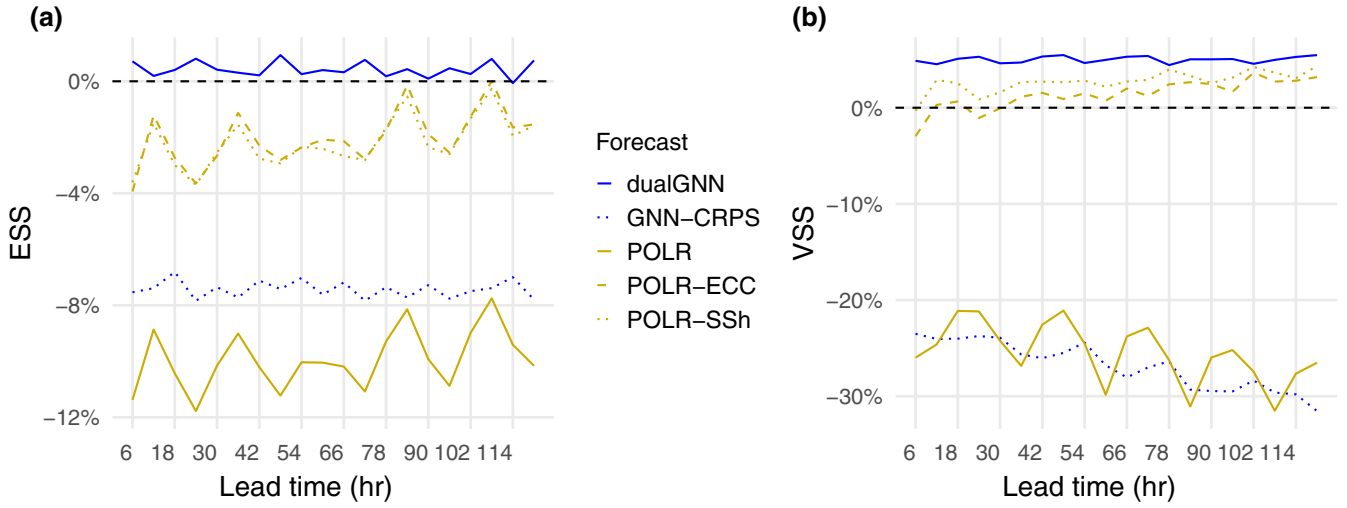

FIGURE 9 (a) Energy skill score (ESS) and (b) variogram skill score (VSS) of proportional odds logistic regression (POLR) and multivariate post-processed visibility forecasts with respect to the GNN-ES forecasts. Both shown as functions of the lead time. CRPS: continuous ranked probability score; ECC: ensemble copula coupling; ES: energy score; GNN: graph neural network; SSh: Schaake shuffle. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 7 Overall mean energy skill score (ESS) and variogram skill score (VSS) of the raw and post-processed visibility forecasts with respect to the energy-score-trained graphical neural network.

	Ensemble	POLR	POLR-ECC	POLR-SSh	GNN-CRPS	dualGNN
ESS (%)	−39.65	−9.95	−2.03	−2.21	−7.43	**0.43**
VSS (%)	−42.51	−25.51	1.37	2.73	−26.83	**5.01**

Abbreviations: CRPS: continuous ranked probability score; ECC: ensemble copula coupling; GNN: graph neural network; POLR: proportional odds logistic regression; SSh: Schaake shuffle.
Note: The value in bold indicates the greatest improvement with respect to the reference.

5.4 | Multivariate performance of visibility forecasts

This section focuses on the multivariate post-processing of visibility forecasts. In addition to the POLR and GNN samples introduced in Section 5.2, we consider the POLR-ECC variant, obtained by reordering the POLR samples according to the rank structure of the raw ECMWF ensemble, and the POLR-SSh models, rearranged based on the rank structure of past observations, as described in Section 5.2. Similar to the previously analyzed variable, the SSh models employ observations from the training period to determine the dependence template.

Figure 9 shows the mean ESS (Figure 9a) and VSS (Figure 9b) of POLR and multivariate post-processed visibility forecasts with respect to the GNN-ES forecasts, as functions of the lead time. The raw forecasts are omitted from the plot to better highlight the differences between the post-processed models. As illustrated in Figure 9a and quantified more precisely in Table 7, the dualGNN shows only a 0.43% improvement, and it is the only model with a positive skill score; nevertheless, the performance of dualGNN and GNN-ES can be considered effectively equivalent. The strongest POLR variant, POLR-ECC, lags slightly behind the reference by 2.03%, closely followed by POLR-SSh, representing a further 0.18% decrease in the mean skill score. The weakest post-processing model is POLR, which exhibits an average 9.95% decrease relative to GNN-ES.

Figure 9b and the second row of Table 7 allows quantification of the improvement of the models relative to the ES-minimizing GNN. On average, the GNN trained to minimize CRPS performs 26.83% worse, whereas the baseline POLR model exhibits a similar decrease of 25.51%. Among the post-processing approaches, POLR-ECC is the first to achieve a positive gain relative to GNN-ES, with an improvement of 1.37%, followed by POLR-SSh with 2.73%. The largest gain is observed for dualGNN, which outperforms GNN-ES by 5.01%. In addition to these results, Supporting Information Figure S9 presents the mean ES and VS for all visibility forecasts.

Figure 10 presents boxplots of DM test statistics, assessing the significance of differences in ES relative to POLR-ECC (the strongest GNN-independent two-step

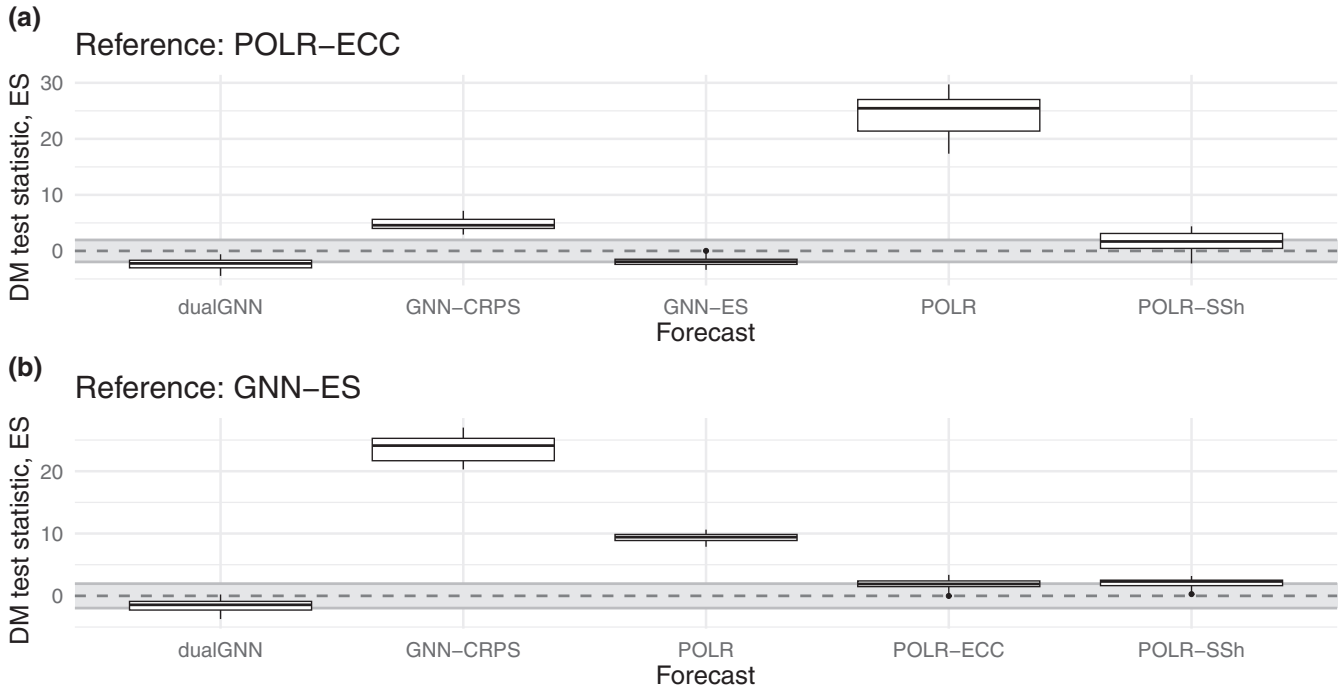

FIGURE 10 Boxplots of Diebold–Mariano (DM) test statistics investigating the significance of the difference in energy score (ES) (a) from the reference POLR-ECC and (b) GNN-ES methods as functions of the lead time. Gray region indicates the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance, and negative values indicate better predictive performance compared with the reference. CRPS: continuous ranked probability score; ECC: ensemble copula coupling; GNN: graph neural network; MLP: multilayer perceptron; POLR: proportional odds logistic regression; SSh: Schaake shuffle.

model) and GNN-ES across different lead times. The gray area marks the acceptance region of the two-tailed DM test at the 5% significance level, and, as in previous analyses, the Benjamini–Hochberg procedure was applied to control the false discovery rate.

As shown in Figure 10a, only GNN-ES and dualGNN significantly outperform the reference POLR-ECC, with improvements observed in 10% and 45% of lead times respectively. For the remaining lead times, the null hypothesis of equal performance cannot be rejected, making these the only models that surpass POLR-ECC. Consistent with the findings from Figure 9 and Table 7, GNN-CRPS and POLR remain below the reference, whereas POLR-SSh achieves comparable performance in 55% of lead times but falls short for the rest.

In Figure 10b, we again adopt the ES-minimizing GNN as the reference model. Relative to this benchmark, only dualGNN shows an improvement, outperforming GNN-ES in 25% of lead times and matching its performance in the remaining cases, making it the only model with consistent gains. POLR-ECC follows, equaling GNN-ES in 90% of lead times and underperforming in the remainder, with POLR-SSh trailing behind. As in previous analyses, GNN-CRPS and POLR never surpass the reference. For a comparison of VS deviations with respect to the reference models, we refer to Supporting Information Figure S11, where the results are briefly discussed.

Finally, the improved calibration of the post-processed models can be further examined in Figure 11, although the histograms appear somewhat more complex compared with Figure 8. Based on the average ranks, dualGNN seems to perform the best, while the ES and dependence histograms suggest a tendency toward overestimated correlation. It is also notable that, in the case of the histograms, the GNN models are not constrained by the maximum generatable value as the POLR model is, which may make them more sensitive to certain patterns.

6 | DISCUSSION

The aim of this study was the multivariate post-processing of ensemble forecasts for different weather quantities, specifically solar irradiance and visibility, using a GNN trained with a combination of the ES and the VS. The primary objective was to demonstrate that an optimal combination of ES and VS can outperform a GNN trained solely on ES, as well as traditional two-step methods based on empirical copulas.

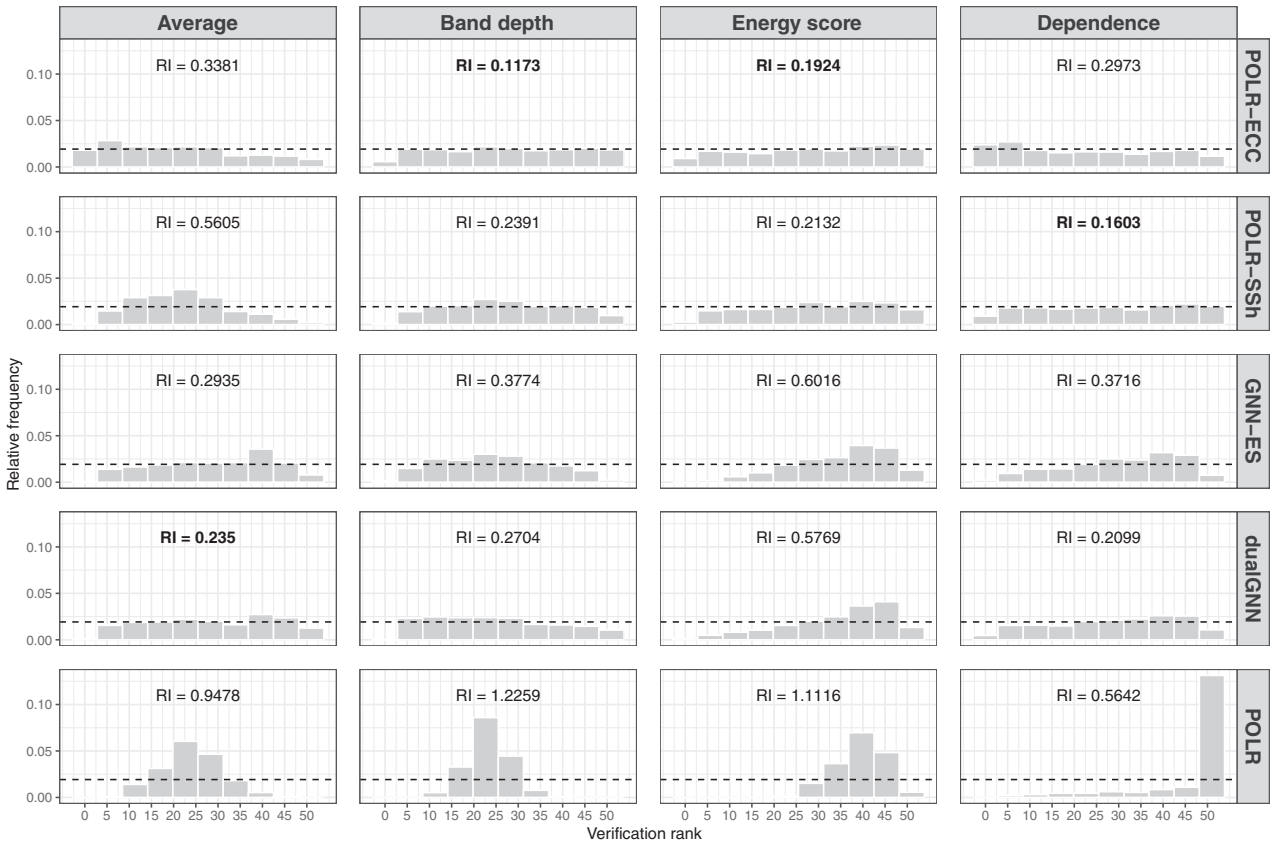

FIGURE 11 Rank histograms of post-processed visibilty forecasts together with the corresponding reliability indices (RI). ECC: ensemble copula coupling; ES: energy score; GNN: graph neural network; POLR: proportional odds logistic regression; SSh: Schaake shuffle.

For both weather variables, the composite-loss GNN (dualGNN) consistently improved upon all reference models, achieving the best performance according to multivariate scores while maintaining well-calibrated marginal distributions. In contrast, empirical copula-based methods did not yield a single clear winner, whereas dualGNN achieved the best performance across both ES and VS. In the case of visibility forecasts, dualGNN also achieved lower mean CRPS than the GNN-ES, surpassed only by a GNN specifically optimized for CRPS. Furthermore, the rank structure of forecasts generated by dualGNN captured relevant dependencies among stations, producing the most uniform rank histograms. Being non-parametric, the method is applicable to any weather variable and allows flexibility in the number of ensemble members generated. Overall, the dualGNN consistently generated accurate forecasts across target variables with diverse characteristics. This demonstrates that the model can robustly handle variables with markedly different statistical properties while sustaining high predictive skill and preserving meaningful rank dependencies.

Despite these advantages, the dualGNN has limitations, as is often the case for many machine-learning methods compared with classical statistical models. One such limitation is increased data requirements, which were particularly evident for visibility forecasts. Though an optimal training period could be identified for solar irradiance with the available data, visibility post-processing would likely benefit from additional training data. However, to maintain a reasonably long verification period, the training period selected here was used in this study.

As with any spatial GNN application, the construction of the graph is crucial. A 50 km threshold for defining edges between stations in northern Chile represents a reasonable choice, although finer graph structures may be more appropriate for visibility forecasts. In the current datasets, clustering stations based on climatological or geographical similarity followed by subsequent edge definition did not provide additional benefits. Nevertheless, a principal advantage of the multivariate loss function lies in the ability of the GNN architecture to produce calibrated samples, even when the underlying graphs are imperfectly specified, while simultaneously capturing dependency structures through joint optimization.

Future research could explore the inclusion of additional variables correlated with the target variable as features, potentially enabling richer feature sets and more informative graph structures, thereby improving

performance. Alternative graph architectures, such as graph attention networks or graph convolutional networks, could also be investigated.

An additional benefit of the GraphSAGE architecture used here is its scalability for large graphs, making it promising for future work on a global network of stations, as well as for comparison with other multivariate score-based models (Chen *et al.*, 2024). Moreover, one of the most promising aspects of dualGNN is its potential for spatial interpolation. However, for variables measured across highly heterogeneous environments, such as the diverse elevations and coastal locations in Chile, a substantially richer graph structure would be required. Preliminary tests suggest that interpolation can yield up to a 20% improvement for certain stations without observations, whereas locations with significantly different climatological characteristics require further investigation. Comparative studies with spatial interpolation models, such as of Baran and Lakatos (2024), would also be of interest.

ACKNOWLEDGEMENTS

I gratefully acknowledge the support of the National Research, Development, and Innovation Office under grant no. K142849, and the EKÖP-25-4-II University Research Scholarship Program of the Ministry for Culture and Innovation, funded by the National Research, Development, and Innovation Fund. Furthermore, I thank my former supervisor, Sándor Baran, for kindly reviewing and providing valuable feedback on the manuscript, and the two anonymous Reviewers for their constructive comments and suggestions.

FUNDING INFORMATION

Hungarian National Research, Development and Innovation Office (grant/award number: K142849) and EKÖP-25-4-II University Research Scholarship Program of the Ministry for Culture and Innovation, funded by the National Research, Development, and Innovation Fund.

DATA AVAILABILITY STATEMENT

The visibility data used in this study are subject to confidentiality restrictions but can be obtained from ECMWF for research purposes. Solar irradiance data are available from the author upon reasonable request.

REFERENCES

Allen, S., Ziegel, J. & Ginsbourger, D. (2024) Assessing the calibration of multivariate probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 150, 1315–1335.

Baran, Á. & Baran, S. (2024) A two-step machine-learning approach to statistical post-processing of weather forecasts for power generation. *Quarterly Journal of the Royal Meteorological Society*, 150, 1029–1047.

Baran, S. & Lakatos, M. (2024) Clustering-based spatial interpolation of parametric postprocessing models. *Weather and Forecasting*, 39, 1591–1604.

Baran, S., Marín, J.C., Cuevas, O., Díaz, M., Szabó, M., Nicolis, O. et al. (2025) Machine-learning-based probabilistic forecasting of solar irradiance in Chile. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11, 89–105.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 57, 289–300.

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, e230027.

Bülte, C., Maskey, S., Scholl, P., von Berg, J. & Kutyniok, G. (2025) Graph neural networks for enhancing ensemble forecasts of extreme rainfall. arXiv preprint arXiv:2504.05471.

Chen, J., Janke, T., Steinke, F. & Lerch, S. (2024) Generative machine learning methods for multivariate ensemble postprocessing. *The Annals of Applied Statistics*, 18, 159–183.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. & Wilby, R. (2004) The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5, 243–262.

Dai, Y. & Hemri, S. (2021) Spatially coherent postprocessing of cloud cover ensemble forecasts. *Monthly Weather Review*, 149, 3923–3937.

Diebold, F.X. & Mariano, R.S. (1995) Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13, 253–263.

Feik, M., Lerch, S. & Stühmer, J. (2024) Graph neural networks and spatial information learning for post-processing ensemble weather forecasts. arXiv preprint arXiv: 2407.11050.

Fey, M. & Lenssen, J.E. (2019) Fast graph representation learning with PyTorch geometric. arXiv preprint arXiv:1903.02428.

Gneiting, T. & Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., Raftery, A.E., Westveld, A.H., III & Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118.

Hamill, T.M. & Colucci, S.J. (1997) Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125, 1312–1327.

Hamilton, W., Ying, Z. & Leskovec, J. (2017) Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 1024–1034.

Jiang, W. & Luo, J. (2022) Graph neural network for traffic forecasting: a survey. *Expert Systems with Applications*, 207, 117921.

Knüppel, M., Krüger, F. & Pohle, M.-O. (2022) Score-based calibration testing for multivariate forecast distributions. arXiv preprint arXiv:2211.16362.

Lakatos, M. & Baran, S. (2024) Enhancing multivariate post-processed visibility predictions utilizing copernicus atmosphere monitoring service forecasts. *Meteorological Applications*, 31, e70015.

Lakatos, M., Lerch, S., Hemri, S. & Baran, S. (2023) Comparison of multivariate post-processing methods using global ECMWF

ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 149, 856–877.

Lerch, S. & Baran, S. (2017) Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 66, 29–51.

Li, W., Pan, B., Xia, J. & Duan, Q. (2022) Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of Hydrology*, 605, 127301.

Li, L., Zhang, Y., Wang, G. & Xia, K. (2025) Kolmogorov–Arnold graph neural networks for molecular property prediction. *Nature Machine Intelligence*, 7, 1346–1354.

McCullagh, P. (1980) Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society. Series B, Methodological*, 42, 243–268.

Möller, A., Lenkoski, A. & Thorarinsdottir, T.L. (2013) Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139, 982–991.

Pic, R., Dombry, C., Naveau, P. & Taillardat, M. (2025) Distributional regression u-nets for the postprocessing of precipitation ensemble forecasts. *Artificial Intelligence for the Earth Systems*, 4, 240067.

Raftery, A.E., Gneiting, T., Balabdaoui, F. & Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174.

Rasp, S. & Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146, 3885–3900.

Schefzik, R., Thorarinsdottir, T.L. & Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28, 616–640.

Scheuerer, M. & Hamill, T.M. (2015) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143, 1321–1334.

Schulz, B., El Ayari, M., Lerch, S. & Baran, S. (2021) Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting. *Solar Energy*, 220, 1016–1031.

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Liu, Z., Berner, J. et al. (2019) A description of the advanced research WRF version 4. NCAR technical note NCAR/TN-556+STR. https://api.semanticscholar.org/CorpusID:196211930 [Accessed: 31st July 2025].

Thorarinsdottir, T.L. & Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society. Series A, Statistics in Society*, 173, 371–388.

Thorarinsdottir, T.L., Scheuerer, M. & Heinz, C. (2016) Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, 25, 105–122.

Vannitsem, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Hemri, S. et al. (2021) Statistical postprocessing for weather forecasts: review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102, E681–E699.

Wilks, D.S. (2019) *Statistical methods in the atmospheric sciences*, 4th edition. Amsterdam: Elsevier.

World Meteorological Organization. (2018) *Guide to instruments and methods of observation. Volume I – Measurement of meteorological variables (WMO-No. 8)*. Geneva: WMO.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lakatos, M. (2026) A composite-loss graph neural network for the multivariate post-processing of ensemble weather forecasts. *Quarterly Journal of the Royal Meteorological Society*, e70119. Available from: https://doi.org/10.1002/qj.70119