DSSR

Discussion Paper No. 158

Dynamic conditional density network of wind vectors

Keigo Sato and Yasumasa Matsuda

October, 2026

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Dynamic conditional density network of wind vectors

Keigo Sato* and Yasumasa Matsuda

Graduate School of Economics and Management, Tohoku University, Japan

Abstract

We propose the dynamic conditional density network (DCDN) to correct the bivariate normal predictive distribution commonly used to postprocess ensemble forecasts of wind vectors. The conditional distribution of the observed wind vector given the ensemble forecast may depart from the bivariate normal distribution. DCDN multiplies a bivariate normal reference density by a correction density. The correction density is a B-spline mixture on the probability integral transform scale of the reference density, and a neural network predicts its weights. DCDN equals its reference density when the correction density is uniform. We apply DCDN to ensemble forecasts of the European Centre for Medium-Range Weather Forecasts and observations at 154 sites in the Tohoku region of Japan. DCDN significantly improved the logarithmic score, energy score, and variogram score over the bivariate normal distribution predicted by a neural network. The improvement in the logarithmic score was large near zero wind speed and at higher wind speeds. This is consistent with non-normal features of the conditional distribution that the bivariate normal distribution misses.

Keywords: wind vector; ensemble postprocessing; distributional regression network; B-spline mixtures; probabilistic forecasting

1 Introduction

Weather forecasts are based on numerical weather prediction (NWP) models. Ensemble prediction systems address the uncertainty of NWP forecasts by generating a set of forecasts, called an ensemble, from perturbed initial conditions or model physics. Ensemble forecasts, however, tend to be biased and underdispersed, in that the ensemble spread is too small to be realistic (Schuhen et al., 2012). Statistical postprocessing corrects these errors by transforming the ensemble into a predictive distribution (Gneiting et al., 2005).

For wind vectors, postprocessing has used the bivariate normal distribution as the predictive distribution. Probabilistic wind forecasts are important for applications such as air traffic control, ship routing, and wind energy (Schuhen et al., 2012). Schuhen et al. (2012) proposed bivariate ensemble model output statistics (BIV-EMOS), and Lang et al. (2019) proposed a distributional regression model, both based on the bivariate normal distribution. Buchner et al. (2026) proposed the bivariate distributional regression neural network (BIV-DRN) with the same distribution.

The bivariate normal assumption does not always match the conditional distribution of the observed wind vector given the forecast. A bivariate normal distribution is fully determined by its mean and covariance, and its density is necessarily unimodal and elliptically symmetric. However, Schuhen et al. (2012) suggested that the conditional distribution of the observed wind vector, given the ensemble forecast, tends to be skewed. The normal margins and the Gaussian dependence implied by the bivariate normal distribution might also not be justified by the data at all sites (Buchner et al., 2026).

In a recent comparison for wind vectors, BIV-DRN had better scores than more flexible methods (Buchner et al., 2026). The flexible methods combine kernel density estimates of the margins with pair copulas, instead of the bivariate normal distribution. Over 60 sites, BIV-DRN had the best logarithmic

*E-mail: sato.keigo.s7@dc.tohoku.ac.jp

score (LogS), energy score (ES), and variogram score (VS) among the compared methods. The flexible methods were nevertheless better calibrated than the methods based on the bivariate normal distribution.

In contrast to these methods, we propose the dynamic conditional density network (DCDN), which corrects a bivariate normal distribution instead of replacing it. DCDN multiplies a bivariate normal reference density, here that of BIV-DRN, by a correction density. The correction density is defined on the probability integral transform (PIT) scale of the reference density. Following Matsuda and Iwafuchi (2025), it is a mixture of tensor products of normalized B-splines. A neural network predicts its mixture weights for each forecast. When the correction density is uniform, the predictive distribution of DCDN equals the bivariate normal reference distribution.

Over the test period of 2025, DCDN significantly improved the LogS, the ES, and the VS over BIV-DRN. We compare DCDN with the raw ensemble forecasts (Raw ensemble), BIV-EMOS, and BIV-DRN. The data are ensemble forecasts of the European Centre for Medium-Range Weather Forecasts and observations at 154 sites in the Tohoku region of Japan. Relative to BIV-EMOS, DCDN improved the LogS by 18.9%, whereas BIV-DRN improved it by 6.0%. In the ES and the VS, the improvements were 3.3% and 4.0% for DCDN and 2.5% and 3.3% for BIV-DRN. The improvement in the LogS was large near zero wind speed and at higher wind speeds.

Section 2 describes the data, and Section 3 the methods. Section 4 presents the results, and Section 5 discusses them and concludes.

2 Data

2.1 Observations

The observations are wind vectors recorded by the Japan Meteorological Agency (JMA) at 154 sites in the Tohoku region of Japan. They are retrieved from the JMA website for downloading past weather data (Japan Meteorological Agency, 2026). Fig. 1 shows the locations of the sites. Each wind vector $\mathbf{y}$ consists of the zonal component y_u and the meridional component y_v.

The observations are recorded as discrete values. Wind speeds are recorded in units of $0.1\,\mathrm{m\,s^{-1}}$, and the nonzero wind vectors point in 16 directions at intervals of 22.5°. JMA records calm, a wind speed of $0.2\,\mathrm{m\,s^{-1}}$ or less, without a direction (Japan Meteorological Agency, 1998), and we treat it as the zero vector. Because the recorded observations are discrete, training and evaluation use dequantized observations.

A dequantized observation $\tilde{\mathbf{y}}_{\mathrm{obs}}$ is drawn uniformly from the set of wind vectors that would be recorded as the observation $\mathbf{y}_{\mathrm{obs}}$. For a non-calm observation, this set is the sector within $0.05\,\mathrm{m\,s^{-1}}$ of its speed and 11.25° of its direction. For a calm observation, it is the disk of radius $0.25\,\mathrm{m\,s^{-1}}$ centered at the origin. This radius adds half the speed interval to the upper limit of calm. The perturbations are drawn anew for each training batch but are fixed for each validation and test case and shared by all methods.

Dequantization prevents a continuous predictive density from raising its likelihood by concentrating on the recorded values. Without it, fitting a continuous density to discrete data produces a degenerate solution that places all probability mass on the discrete values (Ho et al., 2019). In density modeling of images, the discrete pixel values are therefore commonly dequantized by adding uniform noise over the width of each discrete bin (Ho et al., 2019). Schuhen et al. (2012) replaced each wind speed observation of zero by a uniform random number between 0 and 2 kt when computing the continuous ranked probability score. Our disk for calm extends this replacement to wind vectors.

2.2 Ensemble forecasts

The forecasts are ensemble forecasts of the European Centre for Medium-Range Weather Forecasts (ECMWF) from the archive of TIGGE (The International Grand Global Ensemble). TIGGE collects the ensemble predictions of global numerical weather prediction centers (Bougeault et al., 2010; Swinbank et al., 2016). The forecasts are retrieved on the octahedral reduced Gaussian grid O640, whose spacing is about 18 km. They are initialized at 0000 UTC and have 60 lead times from 6 to 360 h at 6-h intervals.

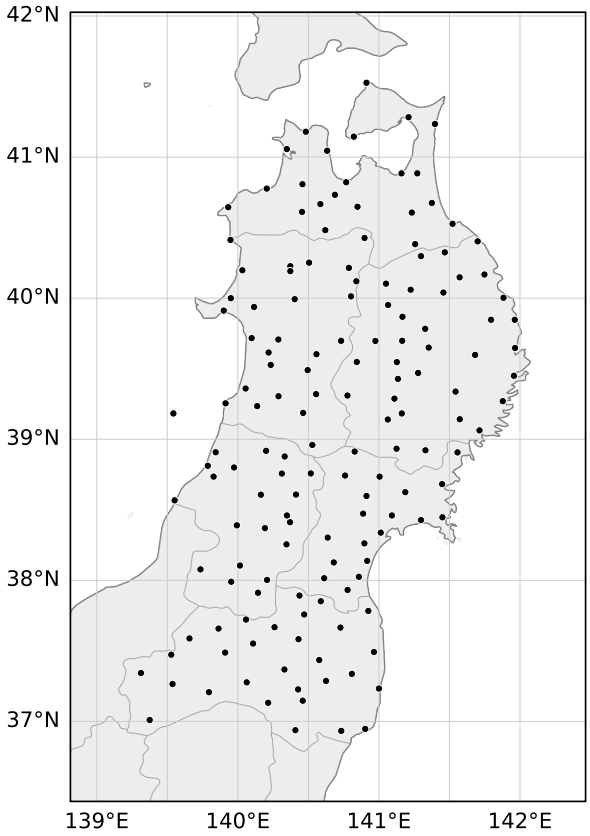


Fig. 1 Locations of the 154 observation sites in the Tohoku region of Japan

Each forecast consists of the control forecast and 50 perturbed members of the zonal and meridional wind components at 10 m.

A case is a combination of a site, an initialization time, and a lead time. For each case, the forecast is taken at the grid point nearest to the site, and the observation at the valid time. Only the cases with an observation are used for training and evaluation.

2.3 Training, validation, and test periods

The data are split into training, validation, and test periods by initialization time. The training period is 2021–2023, the validation period is 2024, and the test period is 2025. The validation period is used for the preliminary search of the hyperparameters and for early stopping, and the test period for the evaluation in Section 4. The initialization times in the last 15 days of each period are excluded. The valid times of the forecasts therefore fall within the period, and adjacent periods are separated by a gap of 15 days. The validation and test periods each have about 3.2 million cases with an observation.

3 Methods

3.1 Predictive distribution of DCDN

The predictive density of DCDN is the reference density g multiplied by a correction density h evaluated at a transformed wind vector (Fig. 2):

$$f(\mathbf{y} \mid \mathbf{x}) = g(\mathbf{y} \mid \mathbf{x})\, h\big(T(\mathbf{y}) \mid \mathbf{x}\big). \tag{1}$$

Here, $\mathbf{x}$ denotes the inputs of a case (Section 3.3). The reference density g is a bivariate normal density with mean $\boldsymbol{\mu}$ and covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_u^2 & \rho\,\sigma_u\sigma_v \\ \rho\,\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}, \tag{2}$$

where σ_u and σ_v are the standard deviations and ρ is the correlation coefficient. In this study, BIV-DRN gives the mean and the covariance from the inputs (Section 3.3).

The transformation T whitens the wind vector with the mean and the symmetric square root of the covariance of g. It then applies the standard normal cumulative distribution function Φ to each component:

$$\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}), \qquad \mathbf{u} = T(\mathbf{y}) = \Phi(\mathbf{z}). \tag{3}$$

If $\mathbf{y}$ follows g, $\mathbf{u}$ is uniformly distributed on the unit square, which we call the PIT scale of the reference density. Because the Jacobian determinant of T equals g, f is a density whenever h is a density on the unit square. In general, h is the density of $T(\mathbf{y})$ when $\mathbf{y}$ follows f.

3.2 Correction density

The correction density h on the PIT scale is built from normalized cubic B-splines, following the representation of densities on the unit hypercube by Matsuda and Iwafuchi (2025). In each coordinate, we use 13 cubic B-splines B_i. Their knots have multiplicity four at 0 and 1 and are spaced 0.1 apart in between. These B-splines are nonnegative and sum to one on the unit interval. The normalized basis functions are

$$\phi_i(u) = \frac{B_i(u)}{m_i}, \qquad m_i = \int_0^1 B_i(u)\,\mathrm{d}u. \tag{4}$$

Each ϕ_i is a density on the unit interval.

The correction density is the mixture

$$h(\mathbf{u} \mid \mathbf{x}) = \sum_{i=1}^{13}\sum_{j=1}^{13} p_{ij}(\mathbf{x})\,\phi_i(u_1)\,\phi_j(u_2), \qquad p_{ij} > 0, \qquad \sum_{i,j} p_{ij} = 1. \tag{5}$$

Because each product of two basis functions integrates to one over the unit square, h is a density on the unit square.

Following Matsuda and Iwafuchi (2025), we transform the correction densities into an unconstrained Euclidean space through generalized logit and softmax mappings of the weights. In this study, we use the weighted softmax

$$p_{ij} = \frac{m_i m_j \exp\big((\mathbf{Q}^\top\mathbf{q})_{ij}\big)}{\sum_{i',j'} m_{i'} m_{j'} \exp\big((\mathbf{Q}^\top\mathbf{q})_{i'j'}\big)}, \tag{6}$$

where $\mathbf{q}$ is a vector of 168 real values. The matrix $\mathbf{Q}$ is a Helmert-type contrast matrix with 168 orthonormal rows of length 169, all orthogonal to the vector of ones. The weights are positive and sum to one, and the mapping is inverted by the logit embedding

$$\mathbf{q} = \mathbf{Q}\log\big(\mathbf{p}/\mathrm{vec}(\mathbf{m}\mathbf{m}^\top)\big), \tag{7}$$

where $\mathbf{p}$ and $\mathbf{m}$ are the vectors of the weights and the integrals, vec follows the order of $\mathbf{p}$, and the operations are componentwise. Weights that are positive and sum to one therefore correspond one to one to $\mathbf{q}$. At $\mathbf{q} = \mathbf{0}$, and only there, each weight equals the product of the integrals of its two basis functions, so h is one everywhere and DCDN equals the reference density.

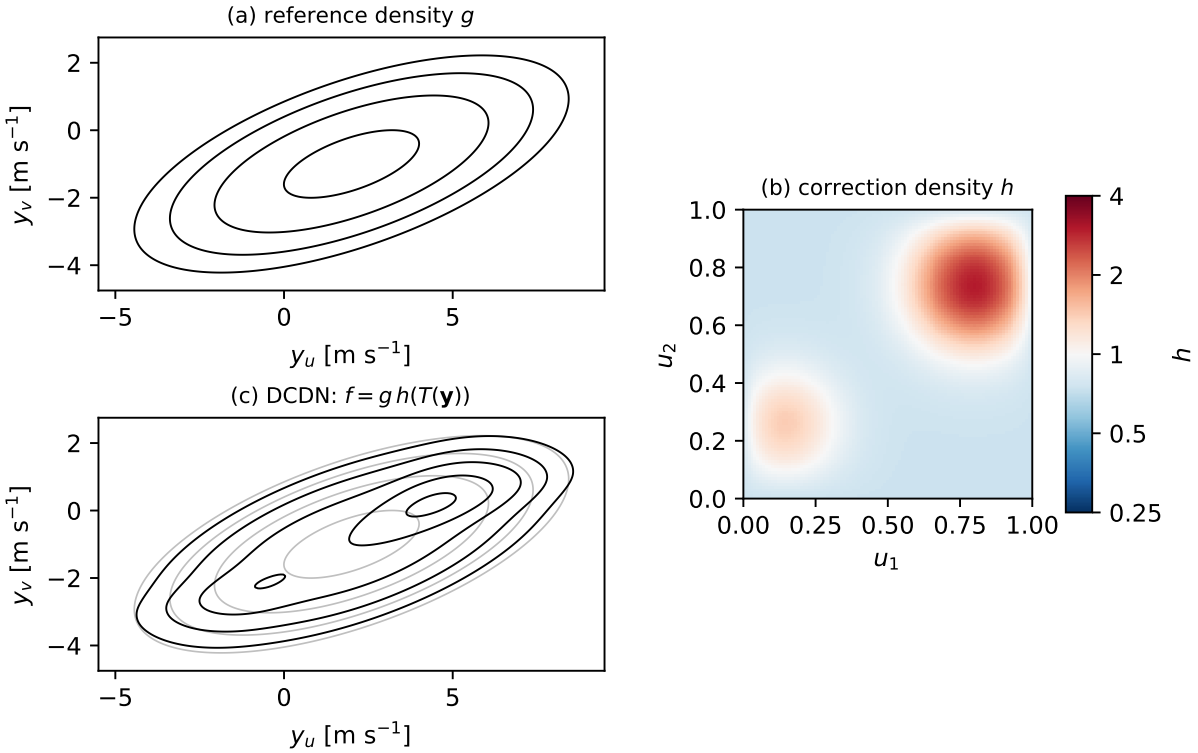


Fig. 2 Schematic of DCDN. (a) Reference density g. (b) Correction density h on the PIT scale. (c) Predictive density f of DCDN, with the contours of g in gray

3.3 Networks and inputs

BIV-DRN and DCDN use the same inputs $\mathbf{x}$: the 16 variables listed in Table 1 (Appendix A.1) and the site. Following Rasp and Lerch (2018), each network represents the site by a site embedding, an entity embedding (Guo and Berkhahn, 2016) that is learned with the weights and concatenated with the other inputs.

BIV-DRN predicts the five parameters of the bivariate normal distribution from the inputs with a fully connected neural network. Buchner et al. (2026) extended the univariate distributional regression network to this bivariate setting for wind vectors. Our BIV-DRN differs from theirs in the inputs and the number of hidden layers, and it is a single network, not an average of ten. Its hidden layers use the rectified linear unit (ReLU) as the activation function, and Section 3.4 gives their number and width. The five outputs are the two means, the logarithms of the two standard deviations, and the transformed correlation coefficient

$$\tilde{\rho} = \frac{\rho}{\sqrt{1-\rho^2}}, \qquad \rho = \frac{\tilde{\rho}}{\sqrt{1+\tilde{\rho}^2}},$$

as in Buchner et al. (2026).

The correction network of DCDN predicts the 168 values $\mathbf{q}$ from the inputs. Like BIV-DRN, it is fully connected with ReLU in the hidden layers (Section 3.4). Its site embedding is separate from that of BIV-DRN. The output layer is initialized at zero, so training of the correction network starts from $\mathbf{q} = \mathbf{0}$.

3.4 Training

Training proceeds in two stages: BIV-DRN is trained first, and the correction network is then trained with BIV-DRN fixed as the reference density. In both stages, the loss is the negative log-likelihood of the dequantized observations. The 16 inputs are standardized with their means and standard deviations over the training period. The networks are implemented in PyTorch, and their weights use its default initialization unless stated otherwise. They are trained with Adam (Kingma and Ba, 2015). Because

Table 1 Input variables of BIV-EMOS, BIV-DRN, and DCDN. BIV-DRN and DCDN also use a site embedding

Input	Number
Control forecast of each component	2
Ensemble mean of each component	2
Logarithm of the ensemble standard deviation of each component	2
Ensemble skewness of each component	2
Ensemble excess kurtosis of each component	2
Ensemble correlation between the two components	1
Lead time	1
Sine and cosine of the time of day	2
Sine and cosine of the day of the year	2
Total	16

the reference density is fixed, any difference between DCDN and BIV-DRN arises from the correction density h alone.

The hyperparameters were fixed on the basis of a preliminary search over the validation period. Both BIV-DRN and the correction network of DCDN have three hidden layers of width 256 and a learning rate of 10^{-3}. The embedding length is 10, and a batch consists of all cases of 32 initialization times. The maximum number of epochs is 200, and the early-stopping patience is 10 epochs. Early stopping monitors the LogS (Section 3.6) over the validation period, and we use the weights from the epoch with the lowest value.

3.5 Compared methods

We compare four methods: Raw ensemble, BIV-EMOS, BIV-DRN, and DCDN. Raw ensemble consists of the 50 perturbed members. BIV-EMOS gives the five outputs of BIV-DRN (Section 3.3) as linear functions of the 16 inputs in Table 1, with site-specific coefficients. These outputs are the two means, the logarithms of the two standard deviations, and the transformed correlation coefficient, as in BIV-DRN. This follows ensemble model output statistics (EMOS), a postprocessing technique based on multiple linear regression (Gneiting et al., 2005). The coefficients start at zero and are trained with the same loss, learning rate, and other training settings as BIV-DRN (Section 3.4).

Our BIV-EMOS differs from that of Schuhen et al. (2012) in the predictors and the estimation method. Its predictors are the 16 inputs rather than the ensemble means, variances, and mean wind direction. Its coefficients are estimated jointly rather than in three phases.

3.6 Scores

We evaluate the forecasts with the LogS, the ES, and the VS, which are proper scoring rules relative to suitable classes of distributions (Gneiting et al., 2008; Scheuerer and Hamill, 2015). The three scores are evaluated at the dequantized observation $\tilde{\mathbf{y}}_{\text{obs}}$ (Section 2.1). All three are oriented as penalties, so a lower score indicates a better forecast.

The LogS uses the predictive density, namely f for DCDN and the bivariate normal densities for BIV-EMOS and BIV-DRN:

$$\text{LogS} = -\ln f(\tilde{\mathbf{y}}_{\text{obs}} \mid \mathbf{x}). \tag{8}$$

This is the logarithmic score (Gneiting et al., 2008) with the natural logarithm.

The ES and the VS use samples $\mathbf{y}_1, \ldots, \mathbf{y}_m$ to approximate expectations under the predictive distribution. For the three postprocessing methods, the samples are $m = 1000$ draws from the predictive distribution (Appendix A.2). For Raw ensemble, the samples are its $m = 50$ members, and only the ES

and the VS are reported. The ES is

$$\mathrm{ES} = \frac{1}{m}\sum_{k=1}^{m}\left\|\mathbf{y}_k - \tilde{\mathbf{y}}_{\mathrm{obs}}\right\| - \frac{1}{2m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}\left\|\mathbf{y}_k - \mathbf{y}_l\right\|, \tag{9}$$

which is the energy score (Gneiting et al., 2008), a multivariate generalization of the continuous ranked probability score, with the Euclidean norm and equal masses on the samples. The VS is

$$\mathrm{VS} = 2\left(|\tilde{y}_{\mathrm{obs},u} - \tilde{y}_{\mathrm{obs},v}|^{0.5} - \frac{1}{m}\sum_{k=1}^{m}\left|y_{k,u} - y_{k,v}\right|^{0.5}\right)^2, \tag{10}$$

which is the variogram score of Scheuerer and Hamill (2015) of order 0.5 with all weights equal to one.

3.7 Comparison and tests

We show the improvements as percentage improvements over BIV-EMOS. Let S_m and S_{EMOS} be a score averaged over all cases for method m and for BIV-EMOS. The improvement of method m is

$$100\left\{\exp(S_{\mathrm{EMOS}} - S_m) - 1\right\} \text{ for the LogS} \quad \text{and} \quad 100\,(1 - S_m/S_{\mathrm{EMOS}}) \text{ for the ES and the VS,}$$

and a positive value means that method m is better than BIV-EMOS. For the ES and the VS, the improvement is the skill score (Gneiting and Raftery, 2007) in percent, computed against BIV-EMOS with an optimal score of zero. In the LogS, the improvement is the percentage by which the geometric mean of the predictive density at the observations exceeds that of BIV-EMOS.

We repeat the training of BIV-EMOS, BIV-DRN, and DCDN ten times with different random numbers (Appendix A.3). The overall scores and improvements are averaged over the ten training runs. Tables 2 and 3 also give their standard deviations. Raw ensemble is not trained, so its ES and VS are single values. The DM tests and the results shown in figures use the model from one of the ten training runs of each method.

For each of the three scores, a two-sided Diebold–Mariano (DM) test compares the three pairs among BIV-EMOS, BIV-DRN, and DCDN. The test uses the differences in the score between two methods, averaged over all sites and lead times for each initialization time. Let d_t be this difference for the tth of the N initialization times, and let $\bar{d}$ be their mean. The test statistic is

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\hat{V}/N}}, \qquad \hat{V} = \hat{\gamma}_0 + 2\sum_{\tau=1}^{L}\hat{\gamma}_\tau, \qquad \hat{\gamma}_\tau = \frac{1}{N}\sum_{t=1}^{N-\tau}(d_t - \bar{d})(d_{t+\tau} - \bar{d}), \tag{11}$$

where $\hat{\gamma}_\tau$ is the sample autocovariance at lag τ, and $\hat{V}$ uses the uniform, or rectangular, lag window of Diebold and Mariano (1995). Because forecasts initialized within 14 days of each other have overlapping valid times, the maximum lag L is 14. The statistic is compared with the standard normal distribution at the significance level of 0.05.

4 Results

4.1 Overall scores

Over the test period of 2025, DCDN had lower LogS, ES, and VS than BIV-DRN (Table 2). Relative to BIV-EMOS, BIV-DRN improved the LogS, the ES, and the VS by 6.00%, 2.48%, and 3.31%, respectively. DCDN improved them by 18.91%, 3.28%, and 4.03%, respectively. All nine DM tests were significant at the 0.05 level, the weakest being the difference between BIV-DRN and DCDN in the VS ($p = 0.043$). Raw ensemble had higher ES and VS than the three postprocessing methods. In each of the ten training runs, the improvement of DCDN exceeded that of BIV-DRN in all three scores.

Table 2 Mean scores and percentage improvements over BIV-EMOS in the test period, given as the mean ± standard deviation over ten training runs. Lower part: DM tests for one training run of each method, where a positive statistic favors the second method of the pair

Method	LogS	ES	VS
Raw ensemble	–	1.7997	1.0928
BIV-EMOS	3.2953 ± 0.0004	1.2813 ± 0.0004	0.5945 ± 0.0002
BIV-DRN	3.2371 ± 0.0069	1.2495 ± 0.0020	0.5748 ± 0.0017
DCDN	3.1222 ± 0.0038	1.2392 ± 0.0010	0.5706 ± 0.0008
Improvement of BIV-DRN (%)	6.00 ± 0.73	2.48 ± 0.17	3.31 ± 0.29
Improvement of DCDN (%)	18.91 ± 0.45	3.28 ± 0.10	4.03 ± 0.13

Score	Pair	DM statistic	p-value
LogS	BIV-EMOS – BIV-DRN	10.33	< 0.001
	BIV-EMOS – DCDN	21.60	< 0.001
	BIV-DRN – DCDN	24.71	< 0.001
ES	BIV-EMOS – BIV-DRN	12.08	< 0.001
	BIV-EMOS – DCDN	15.97	< 0.001
	BIV-DRN – DCDN	7.17	< 0.001
VS	BIV-EMOS – BIV-DRN	12.67	< 0.001
	BIV-EMOS – DCDN	9.20	< 0.001
	BIV-DRN – DCDN	2.03	0.043

The calm cases accounted for about a fifth of the differences in the mean LogS and ES between BIV-DRN and DCDN (Table 3). In the VS, they accounted for more than the whole difference. They were only 3.21% of the test cases. In the non-calm cases, BIV-DRN improved the LogS, the ES, and the VS over BIV-EMOS by 5.54%, 2.45%, and 3.19%, respectively. DCDN improved them by 16.13%, 3.11%, and 3.03%, respectively. Because the subsets are defined by the observations, ranking the methods within a subset ought to be avoided (Lerch et al., 2017).

4.2 Sites and lead times

The improvement in LogS of DCDN over BIV-DRN is shown for each site in Fig. 3 and for each lead time in Fig. 4. This improvement is the LogS of BIV-DRN minus that of DCDN, averaged over the cases at each site or lead time. Each improvement is shown for all cases and for the non-calm cases.

The improvement was positive at 144 of the 154 sites in all cases and at 139 sites in the non-calm cases (Fig. 3). In all cases, the median improvement across the sites was 0.108, and the improvement ranged from −0.108 to 0.365. In the non-calm cases, the median was 0.086, and the improvement ranged from −0.141 to 0.256. In the non-calm cases, many of the sites with the largest improvements were near the coast of the Sea of Japan. In both sets of cases, the sites with negative improvements were scattered.

The improvement increased with lead time in all cases and in the non-calm cases (Fig. 4). In all cases, it was positive at all 60 lead times and increased from 0.060 at 6 h to 0.143 at 360 h. In the non-calm cases, it was positive at all lead times except 18 h (−0.0003) and increased from 0.059 at 6 h to 0.132 at 360 h. The daily oscillation of the improvement was larger in all cases than in the non-calm cases.

4.3 Observed wind speeds

The improvement in LogS of DCDN over BIV-DRN in the non-calm cases is shown for each interval of the recorded wind speed in Fig. 5. The non-calm cases are grouped into intervals of $1\,\mathrm{m\,s^{-1}}$, with the cases of $15\,\mathrm{m\,s^{-1}}$ or more merged. As in Fig. 3, the improvement is the LogS of BIV-DRN minus that of DCDN averaged over the cases in each interval.

Table 3 As the upper part of Table 2, separately for non-calm and calm cases

Method	LogS	ES	VS
Non-calm cases (3,121,733 cases, 96.79%)			
Raw ensemble	–	1.8241	1.0692
BIV-EMOS	3.3595 ± 0.0004	1.3088 ± 0.0005	0.5891 ± 0.0002
BIV-DRN	3.3055 ± 0.0076	1.2767 ± 0.0021	0.5703 ± 0.0018
DCDN	3.2100 ± 0.0045	1.2682 ± 0.0010	0.5713 ± 0.0011
Improvement of BIV-DRN (%)	5.54 ± 0.80	2.45 ± 0.17	3.19 ± 0.32
Improvement of DCDN (%)	16.13 ± 0.52	3.11 ± 0.10	3.03 ± 0.18
Calm cases (103,484 cases, 3.21%)			
Raw ensemble	–	1.0619	1.8041
BIV-EMOS	1.3607 ± 0.0019	0.4493 ± 0.0003	0.7581 ± 0.0009
BIV-DRN	1.1734 ± 0.0227	0.4283 ± 0.0070	0.7109 ± 0.0201
DCDN	0.4742 ± 0.0263	0.3661 ± 0.0046	0.5493 ± 0.0094
Improvement of BIV-DRN (%)	20.63 ± 2.70	4.68 ± 1.52	6.23 ± 2.58
Improvement of DCDN (%)	142.74 ± 6.55	18.51 ± 1.05	27.55 ± 1.28

The improvement was large near zero wind speed and above about $5\,\mathrm{m\,s^{-1}}$, and small between 1 and $3\,\mathrm{m\,s^{-1}}$. The mean improvement was 0.699 in the calm cases (Table 3) and 0.185 in the non-calm cases below $1\,\mathrm{m\,s^{-1}}$ (Fig. 5). Between 1 and $3\,\mathrm{m\,s^{-1}}$, it was about 0.03, the smallest of all intervals. Above $3\,\mathrm{m\,s^{-1}}$, it increased with wind speed and was between 0.13 and 0.23 in every interval from $5\,\mathrm{m\,s^{-1}}$ upward. As with Table 3, this breakdown by the observations is descriptive.

4.4 Transformed observations

Fig. 6 compares the two-dimensional PIT histogram of the transformed observations with the uniform density and the average correction density. The transformed observation $\mathbf{u}_{\mathrm{obs}}$ is the dequantized observation mapped by T of BIV-DRN, and the histogram is normalized as a density. If the observations follow BIV-DRN, $\mathbf{u}_{\mathrm{obs}}$ is uniformly distributed on the PIT scale. If they follow DCDN, the density of $\mathbf{u}_{\mathrm{obs}}$ in each case is the correction density h of that case (Section 3.1). All cases together then follow the mixture of these densities, which equals the average $\bar{h}$ of h over all test cases.

The histogram of the transformed observations departed from the uniform density, and the average correction density showed the same main features (Fig. 6). Both exceeded one in the central part of the PIT scale, with the peak shifted toward small u_1. Both were below one in a band along the edges and exceeded one at the four corners.

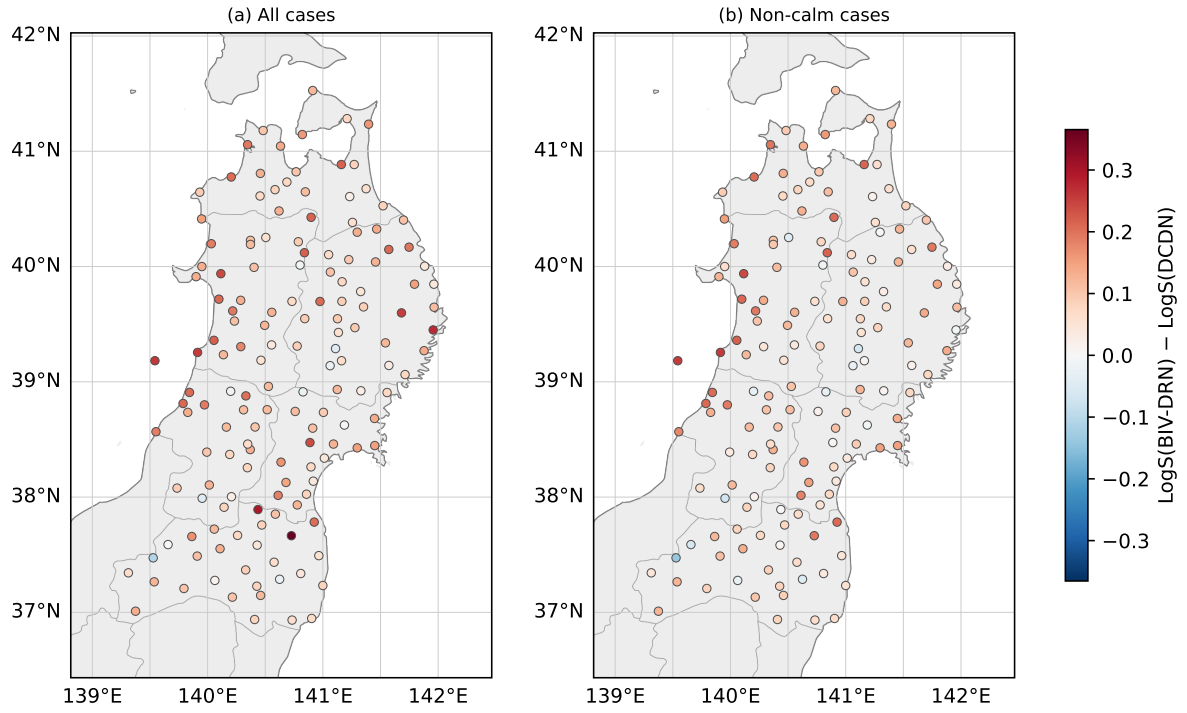


Fig. 3 Improvement of DCDN over BIV-DRN at each site, given as the LogS of BIV-DRN minus that of DCDN averaged over (a) all test cases and (b) non-calm test cases. Positive values favor DCDN

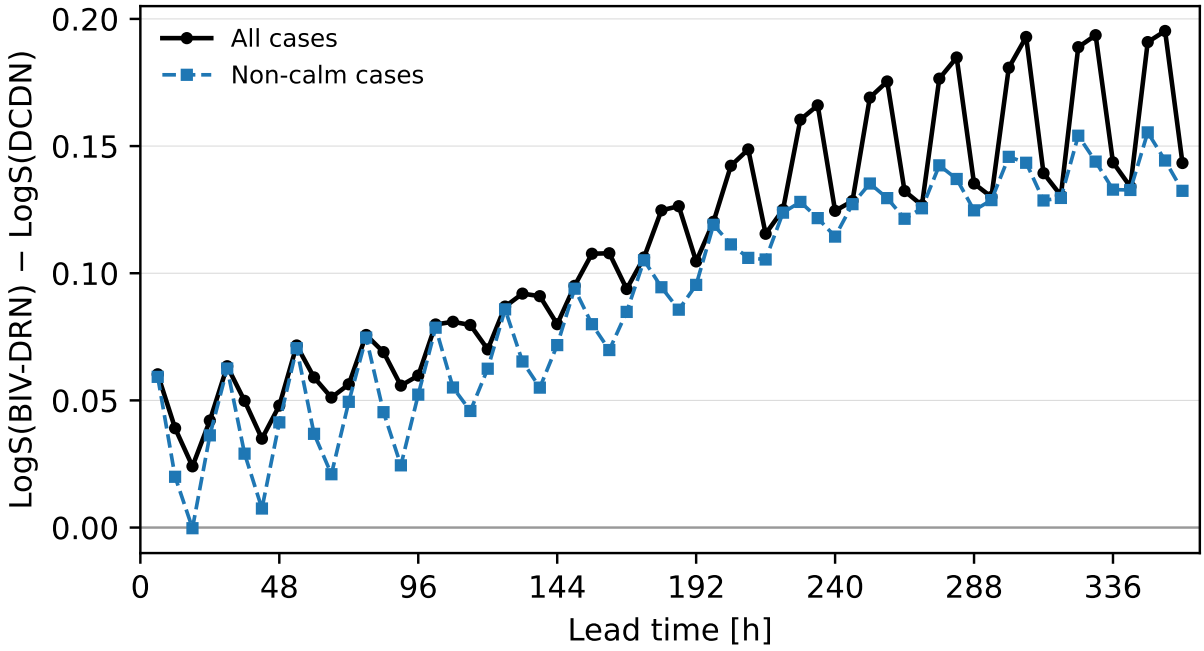


Fig. 4 As Fig. 3, but for each lead time, for all test cases (solid) and non-calm test cases (dashed)

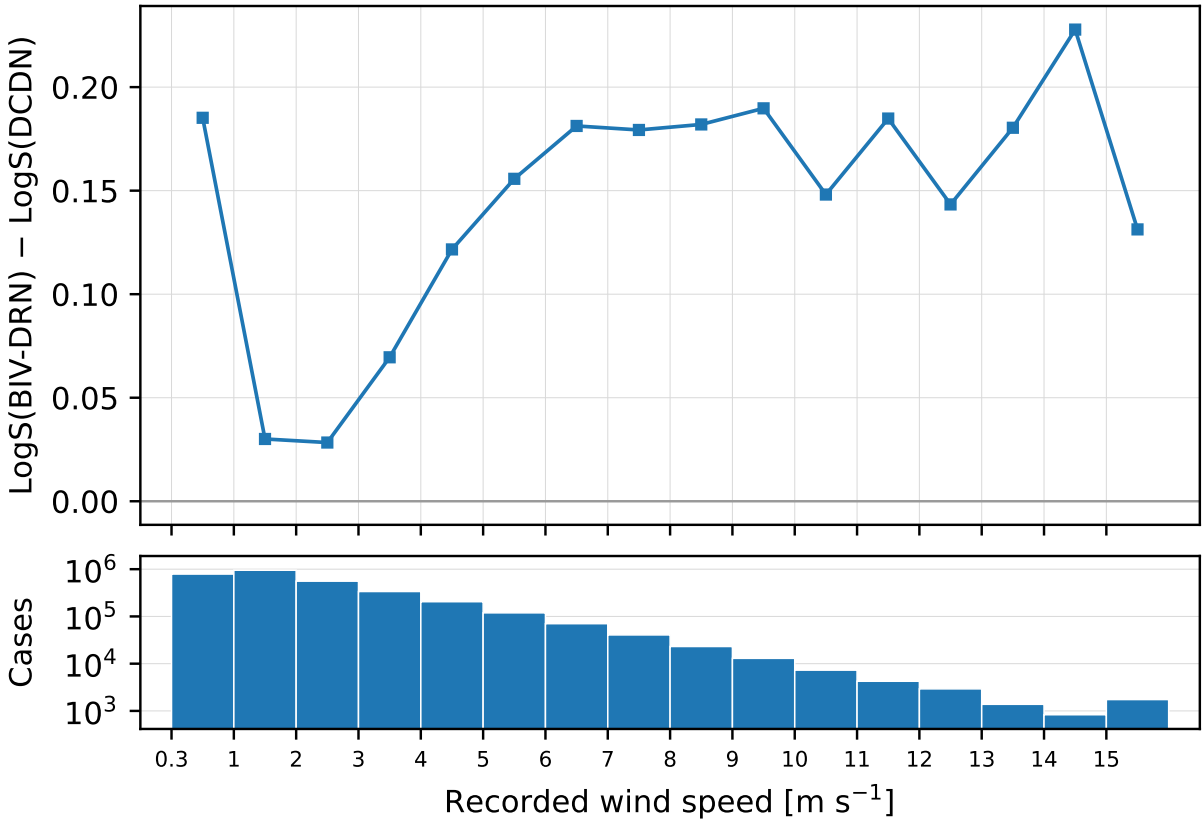


Fig. 5 As Fig. 3, but for each interval of the recorded wind speed of the non-calm test cases. The last interval contains all cases of 15 m s^{-1} or more. Lower panel: number of cases

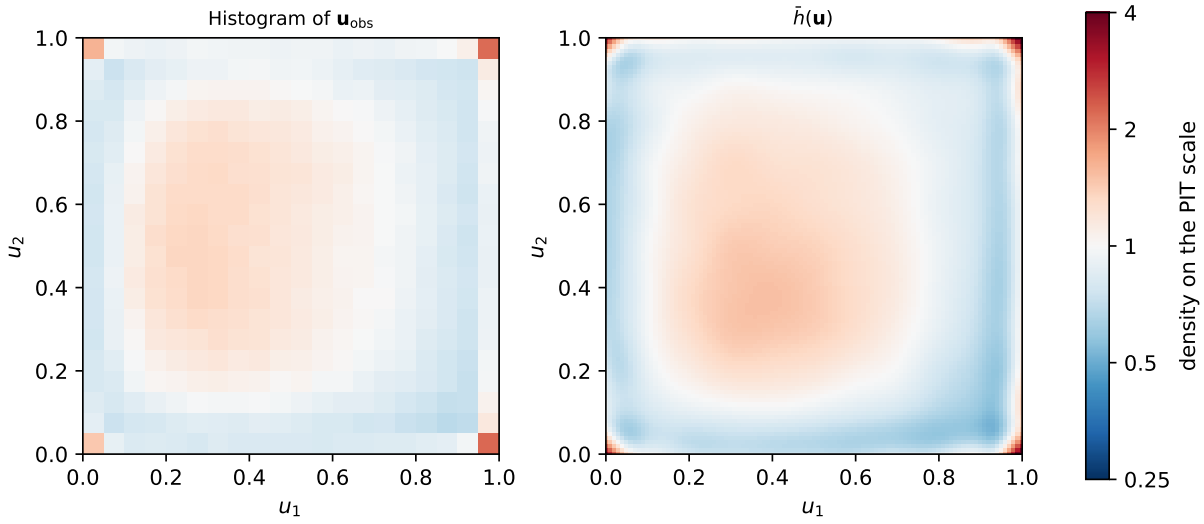


Fig. 6 Test period on the PIT scale. Left: histogram of the transformed observations, normalized as a density. Right: average correction density $\bar{h}$ of DCDN. Values above 4 are shown in the darkest red

5 Discussion and conclusions

DCDN improved on BIV-DRN in all three scores over the test period at 154 sites in the Tohoku region. Relative to BIV-EMOS, DCDN improved the LogS, the ES, and the VS by 18.9%, 3.3%, and 4.0%, respectively. BIV-DRN improved them by 6.0%, 2.5%, and 3.3%. All nine DM tests were significant. The improvement in LogS was positive at most sites and lead times, also in the non-calm cases.

The results are consistent with the interpretation that the bivariate normal distribution of BIV-DRN misses part of the conditional distribution of the observed wind vector. The transformed observations in Fig. 6 departed from the uniform density expected under BIV-DRN, and the average correction density had the same features. With 8 instead of 13 B-splines in each coordinate, DCDN was slightly worse in all three scores, whereas with 23 it improved further only in the LogS. In another variant, the correction was locally restricted to directions whose score functions at zero correction are Fisher orthogonal to those of the reference mean and covariance. This variant retained 96%, 94%, and 81% of the improvement of DCDN over BIV-DRN in the LogS, ES, and VS, respectively.

The main features of the transformed observations are also consistent with specific non-normal features of the conditional distribution. The central peak and the high corners mean that, compared with BIV-DRN, more observations fell near its mean and where both whitened components were extreme.

The improvement in LogS by observed wind speed is also consistent with a conditional distribution concentrated near zero with a relatively heavy tail toward higher wind speeds. The improvement was large near zero wind speed and above about $5\,\mathrm{m\,s^{-1}}$, and small between 1 and $3\,\mathrm{m\,s^{-1}}$ (Section 4.3). For each case, a positive improvement means that DCDN gave the observation a higher density than BIV-DRN, because DCDN multiplies the BIV-DRN density by the correction density. Near zero and at higher wind speeds, DCDN thus gave the observations a higher log density than BIV-DRN on average. The improvement at higher observed wind speeds may be relevant to applications such as wind energy (Schuhen et al., 2012).

The results are limited by the compared methods, the data, and an unverified dequantization assumption. Only BIV-DRN was tested as the reference density, and the comparison excludes the flexible methods of Buchner et al. (2026). The evaluation covers a single region and a single year, 2025. The ECMWF ensemble prediction system was updated several times from the training period to the test period. The dequantization assumes that the wind vector is uniform within the set sharing each recorded value, which the discrete records cannot confirm (Section 2.1). The improvement in the VS is the least certain: it came from the calm cases, and its DM test was not adjusted for the nine comparisons.

DCDN can correct a bivariate normal reference density while containing it as a special case. Because the reference density is kept fixed, the correction can be trained on top of an existing BIV-DRN without retraining it. The correction improved the LogS not only near zero wind speed but also at higher observed wind speeds. This is consistent with non-normal features that the bivariate normal distribution misses. Correcting the bivariate normal distribution, rather than replacing it, can thus add flexibility while improving the scores.

Appendix A Implementation details

A.1 Inputs

The ensemble statistics in Table 1 are computed from the $K = 50$ perturbed members $e_{k,c}$ of each wind component $c \in \{u, v\}$, not from the control forecast. For each component, they are the mean $\bar{e}_c$, the standard deviation s_c, the skewness γ_c, and the excess kurtosis κ_c:

$$\bar{e}_c = \frac{1}{K}\sum_{k=1}^{K} e_{k,c}, \qquad s_c = \left(\frac{1}{K-1}\sum_{k=1}^{K}(e_{k,c} - \bar{e}_c)^2\right)^{1/2},$$

$$\gamma_c = \frac{m_{3,c}}{m_{2,c}^{3/2}}, \qquad \kappa_c = \frac{m_{4,c}}{m_{2,c}^{2}} - 3, \qquad m_{n,c} = \frac{1}{K}\sum_{k=1}^{K}(e_{k,c} - \bar{e}_c)^n.$$

The input for the standard deviation is $\ln s_c$. The correlation between the two components is the sample correlation coefficient

$$r = \frac{\sum_{k=1}^{K}(e_{k,u} - \bar{e}_u)(e_{k,v} - \bar{e}_v)}{\left(\sum_{k=1}^{K}(e_{k,u} - \bar{e}_u)^2 \sum_{k=1}^{K}(e_{k,v} - \bar{e}_v)^2\right)^{1/2}}.$$

The remaining inputs are the control forecast, the lead time, and the time of day and the day of the year. The control forecast of each component is used as it is, and the lead time is given in hours. The time of day and the day of the year are taken at the valid time. Let t be the time of day in hours UTC, and let d be the day of the year, with $d = 1$ on 1 January. Their inputs are

$$\sin\frac{2\pi t}{24}, \qquad \cos\frac{2\pi t}{24}, \qquad \sin\frac{2\pi d}{365.25}, \qquad \cos\frac{2\pi d}{365.25}.$$

A.2 Sampling from DCDN

Samples from DCDN are drawn by choosing a mixture component, drawing each coordinate from its normalized B-spline, and inverting T. First, a component (i, j) is chosen with probability p_{ij}. Second, u_1 and u_2 are drawn independently as averages of the five consecutive knots of B_i and B_j, respectively, with weights drawn uniformly from the simplex. The average of the knots of B_i with these weights has the density ϕ_i (Karlin et al., 1986). Finally, T is inverted to give the sample

$$\mathbf{z} = \Phi^{-1}(\mathbf{u}), \qquad \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}.$$

The sample $\mathbf{y}$ therefore follows f.

A.3 Training and random numbers

The random seed sets the initial weights, the order of the training batches, and the perturbations of the training observations (Section 2.1). It also sets the random numbers for the 1000 samples of the ES and the VS. The ten training runs of each method use the seeds 0 to 9. The perturbations of the validation and test observations use a fixed random seed of 0 for all training runs. Training and evaluation use deterministic computations, so that differences between seeds arise only from these random numbers. The standard deviations of the ES and the VS in Tables 2 and 3 therefore also include sampling variation.

Models of different methods are paired by seed. DCDN with each seed uses BIV-DRN with the same seed as its reference density, and each improvement pairs the models with the same seed. The preliminary search for the hyperparameters used seed 0, and the fixed hyperparameters are used for all seeds. The DM tests and the results shown in figures use the models with seed 0.

References

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.-Y., Parsons, D., Raoult, B., Schuster, D., Silva Dias, P., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., and Worley, S. (2010). The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society*, 91(8):1059–1072.

Buchner, F., Jobst, D., Möller, A., and Czado, C. (2026). Bivariate postprocessing of wind vectors. *Quarterly Journal of the Royal Meteorological Society*, 152(778):e70188.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

European Centre for Medium-Range Weather Forecasts (2026). TIGGE global medium-range ensemble forecasts from multiple NWP centres since October 2006. ECMWF Data Store. `https://doi.org/10.21957/8511726f`. Accessed 26 July 2026.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235.

Guo, C. and Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. (2019). Flow++: Improving flow-based generative models with variational dequantization and architecture design. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2722–2730. PMLR.

Japan Meteorological Agency (1998). Guidelines for meteorological observations. `https://www.jma.go.jp/jma/kishou/know/kansoku_guide/tebiki.pdf`. Accessed 25 October 2026.

Japan Meteorological Agency (2026). Past weather data download. `https://www.data.jma.go.jp/risk/obsdl/index.php`. Accessed 27 July 2026.

Karlin, S., Micchelli, C. A., and Rinott, Y. (1986). Multivariate splines: A probabilistic perspective. *Journal of Multivariate Analysis*, 20(1):69–90.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Lang, M. N., Mayr, G. J., Stauffer, R., and Zeileis, A. (2019). Bivariate Gaussian models for wind vectors in a distributional regression framework. *Advances in Statistical Climatology, Meteorology and Oceanography*, 5(2):115–132.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1):106–127.

Matsuda, Y. and Iwafuchi, R. (2025). Density-valued ARMA models by spline mixtures. *Journal of Time Series Analysis*.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900.

Scheuerer, M. and Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Monthly Weather Review*, 140(10):3204–3219.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., Keller, J. H., Matsueda, M., Methven, J., Pappenberger, F., Scheuerer, M., Titley, H. A., Wilson, L., and Yamaguchi, M. (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, 97(1):49–67.

Acknowledgments

The first author was supported by the Morinokuni Scholarship Foundation and the Shigemitsu Sunaga Fund at Tohoku University.

This work is based on TIGGE data. TIGGE (The International Grand Global Ensemble) is an initiative of the World Weather Research Programme (WWRP).

The first author used Claude Code (Anthropic) and Codex (OpenAI) to write the code and draft the manuscript. The first author reviewed the generated code and text and takes full responsibility for the content of this paper.

Data Availability

The ensemble forecasts are based on data and products of the European Centre for Medium-Range Weather Forecasts (ECMWF), obtained from the TIGGE dataset (ECMWF, 2026; Bougeault et al., 2010; Swinbank et al., 2016). © 2026 ECMWF. Source: ECMWF Data Store at ecds.ecmwf.int. This data is published under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence (`https://creativecommons.org/licenses/by/4.0/legalcode`). ECMWF does not accept any liability whatsoever for any error or omission in the data, their availability, or for any loss or damage arising from their use. We extracted the forecasts at the grid points nearest to the sites.

The wind observations are based on "Past weather data download" (Japan Meteorological Agency) (`https://www.data.jma.go.jp/risk/obsdl/index.php`).