Supplement to "Generative machine learning methods for multivariate ensemble post-processing"

Jieyu Chen[1], Tim Janke[2], Florian Steinke[2] and Sebastian Lerch[1,3]

[1]Karlsruhe Institute of Technology
[2]Technical University of Darmstadt
[3]Heidelberg Institute for Theoretical Studies

September 23, 2022

1 Ablation studies

We here present several ablation studies to illustrate the effects of different choices regarding the architecture and hyperparameters of our conditional generative model (CGM).

1.1 Model hyperparameters

Based on the hyperparameter optimization approach described in the main paper, our CGM implementation utilizes a multivariate normal latent distribution to generate samples of $D_{\text{latent}} = 10$ dimensional samples. The model estimation is based on a batch size of 64, a learning rate of 0.001, and utilizes early stopping with a maximum of 300 epochs and a patience of 10.

In the following, we present several ablation experiments to investigate different choices of hyperparameters. Thereby, we restrict our attention to $D = 10$ and consider a single hyperparameter only, while the rest is kept fixed at the default choice from the main paper. All boxplots shown in the following summarize mean scores over the 100 repetitions of the sampling procedure over the test set (calendar year 2016), and boxplots in red color indicate the results from our CGM implementation.

1.1.1 Latent distribution

Figure 1 illustrates the differences between a normal and uniform latent distribution. For both target variables and the two multivariate scores, the differences are very minor.

1.1.2 Latent dimensions

The latent dimension, i.e., the number of latent variables has a considerable effect on the computational costs. However, as illustrated in Figure 2, different latent dimensions lead to very similar results in terms of the predictive performance of the corresponding generative models.

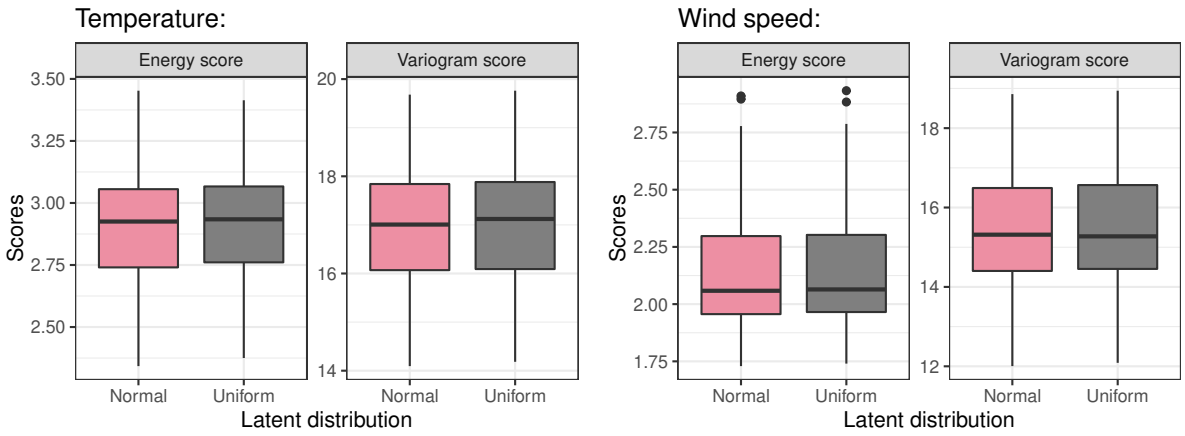

Figure 1: Boxplots of mean scores of different choices for the latent distribution.

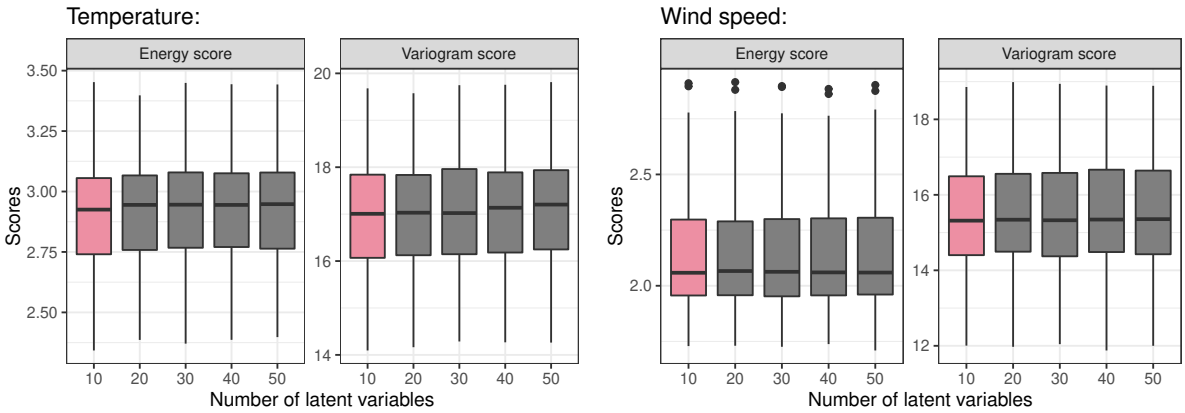

Figure 2: Boxplots of mean scores for different numbers of latent variables.

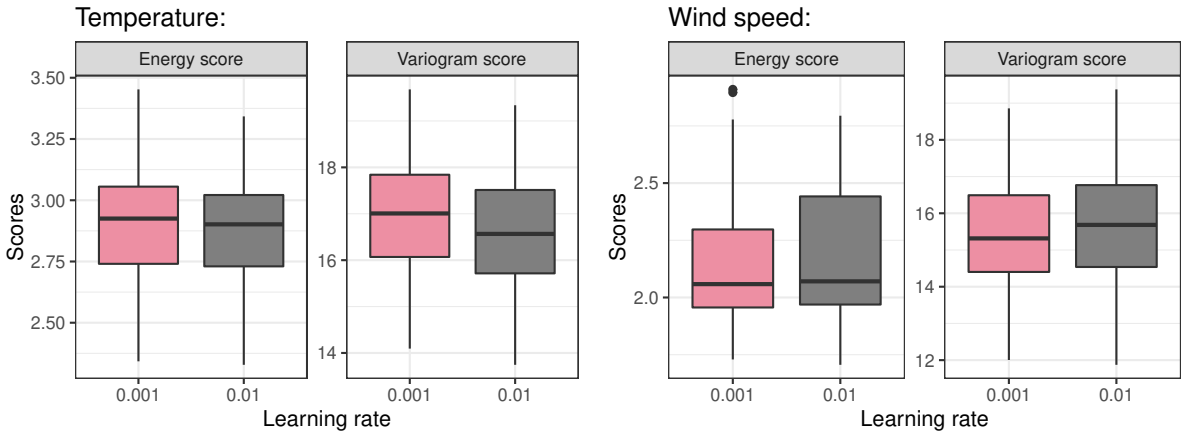

Figure 3: Boxplots of mean scores of different learning rates during model training.

1.1.3 The effects of learning rate

Different choices of the learning rate lead to opposite effects for temperature and wind speed. Figure 3 illustrates that a larger learning rate results in slightly better temperature forecasts in terms of both multivariate scores, but a worse performance for wind speed forecasts. During our experiments, we further noted that a larger learning rate can sometimes lead to unstable model training for wind speed post-processing, likely due to the non-negativity constraint.

1.1.4 The effects of normalized input target variable

Normalizing inputs to neural network (NN)-based models is generally recommended as a standard pre-processing step to stabilize and accelerate the training process. As described in the main paper, we normalize all meteorological weather variables except for the target variable in the first two CGM modules, according to our initial experiments. Figure 4 formally evaluates this choice based on the test data and confirms that normalizing the target variable forecasts has a negative impact on the forecast performance, in particular for wind speed.

1.1.5 The effects of early stopping during training

We employ early stopping to avoid overfitting during training. An alternative strategy would be to train for a fixed number of epochs without early stopping. Different choices are compared in Figure 5. For both temperature and wind speed, the use of early stopping results in comparable or better forecasts with the ones trained for a fixed number of epochs. In practice, our CGM model generally stops training after around 20 epochs when applying early stopping.

1.2 CGM architecture choices

1.2.1 The effects of a nonlinear model for the mean module

As illustrated in the schematic illustration in the Figure 2 of the main paper, we employ a linear model for the mean module of CGM, which implies that the mean component of the output samples is linearly dependent on the means of each meteorological input variable, where the relations (weights) are independent across dimensions. An alternative choice is given by a nonlinear model that utilizes fully connected dense layers for the mean module, similar to the noise decoder module of CGM. The nonlinear model enables the mean component of the output samples at each dimension to depend on the mean values of each meteorological variable in all dimensions. Figure 6 compares these two choices, where we utilize one dense layer with a linear activation function for temperature forecasts, and three dense layers (including two hidden layers consisting of 100 nodes) with an 'elu' activation for wind speed. While results are comparable overall, the nonlinear models lead to slightly better multivariate forecasts.

2 Additional results

Here, we present additional results, e.g., results for other values of D not shown in the main paper, as well as alternative multivariate evaluation metrics. The general experimental setup follows our description in the main paper throughout.

2.1 Univariate results

Figure 7 presents univariate results analogous to those discussed in Section 5.3 of the main paper, but for $D = 10$ and $D = 20$. While there are minor differences due to random effects in the choices of stations, the results are overall similar, indicating that our CGM approach

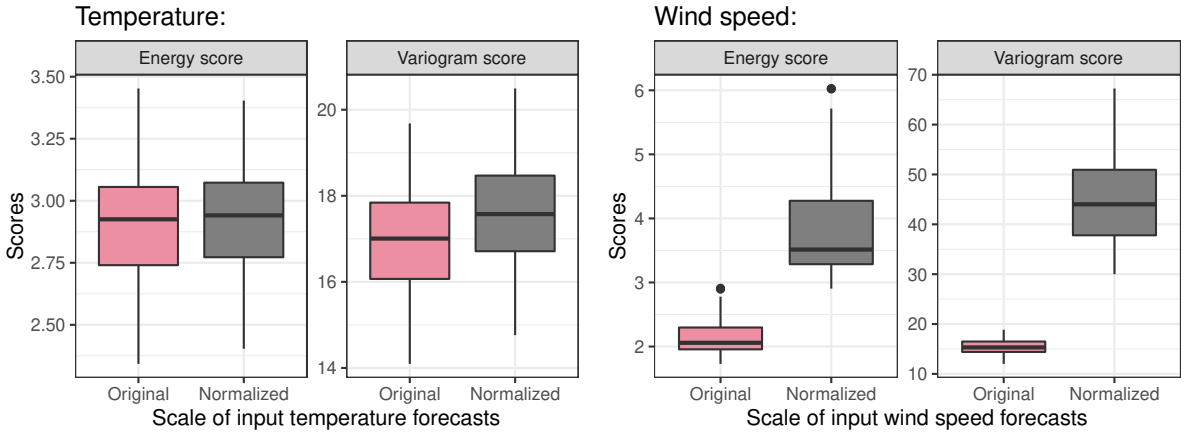

Figure 4: Boxplots of mean scores comparing normalized and non-normalized inputs of target variable forecasts.

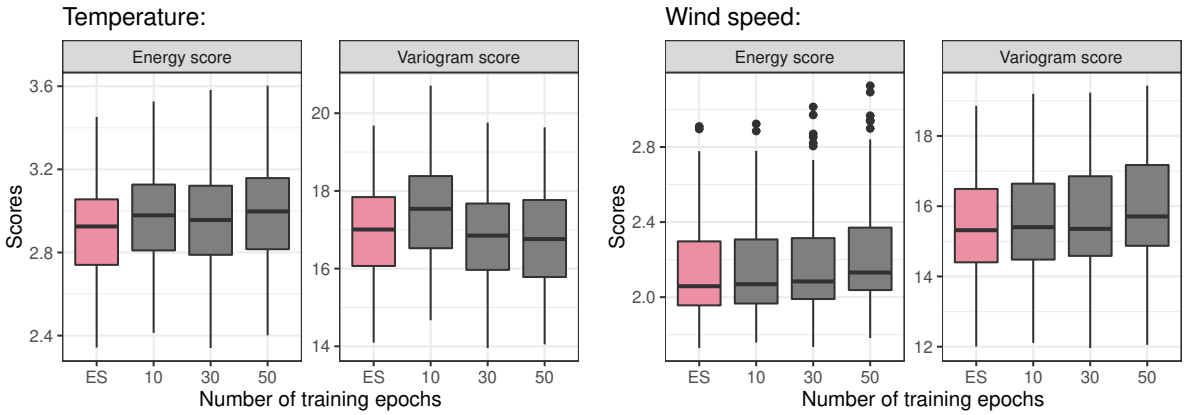

Figure 5: Boxplots of mean scores comparing the use of early stopping or training for a fixed number of epochs.

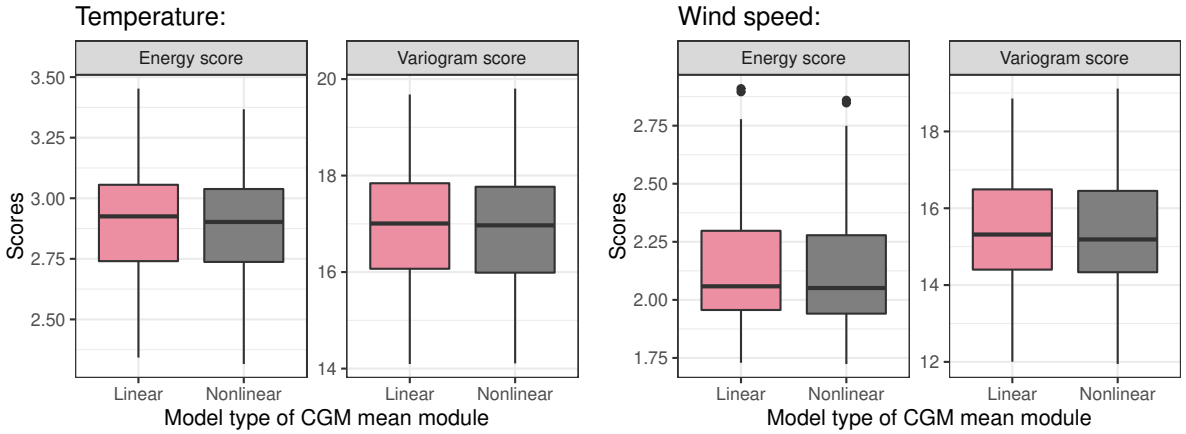

Figure 6: Boxplots of mean scores for different variants of the CGM mean module.

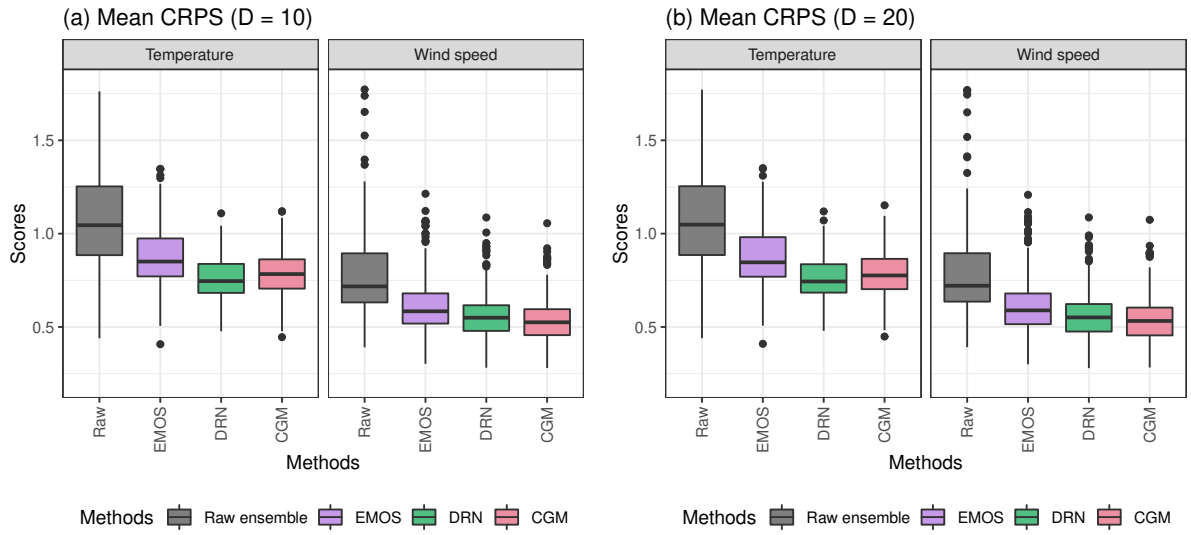

Figure 7: Boxplots of mean CRPS values of different multivariate post-processing methods with $D = 10$ and $D = 20$, including the scores of raw ensemble forecasts, analogous to Figure 3 in the main paper.

also provides skillful univariate forecasts, even when trained for higher-dimensional multivariate settings.

Verification rank and probability integral transform histograms are widely used tools to assess the calibration of univariate forecasts, see, e.g., Thorarinsdottir et al. (2016) for details. Verification rank histograms for the raw and post-processed forecasts are shown in Figures 8 and 9. In general, all post-processing methods notably improve the calibration of the under-dispersed ensemble forecasts. Nevertheless, we observe some minor deviations from uniformity in the rank histograms of the post-processed forecasts, most notably in the case of wind speed where the CGM forecasts show clear improvements over EMOS and DRN.

2.2 Multivariate results

2.2.1 Energy score and variogram score

Figure 10 provides additional results on the effect of the CGM sample size for $D = 5$ and $D = 20$. The results are generally very similar to the case of $D = 10$ covered in the main paper, and we do not observe any structural differences.

Tables 1, 2, 3 and 4 show the rejection rates of DM tests of equal predictive performance to quantify the statistical significance of the multivariate score differences between different post-processing methods across the repetitions of the experiments, here for $D = 5$ and $D = 20$. As before, we do not observe any substantial differences across the choices of D.

In addition to the statistical significance of the observed score differences between different multivariate post-processing methods, we perform similar tests to compare the CGM forecasts based on different number of samples generated from the post-processed multivariate distribution. The corresponding rejection rates of DM test of equal predictive performance are shown in Tables 5, 6, 7, 8, 9 and 10 for all choices of D, and for both target variables and multivariate scores. Generating a larger number of samples from the CGM post-processed distribution generally leads to significant improvements over the default setup with 50 samples, in particular in higher-dimensional settings. As expected, the significance of the score differences diminishes when comparing to a larger number of CGM samples, for example the score differences between 500 and 400 sample-based CGM variants rarely are significant in only a small proportion of the

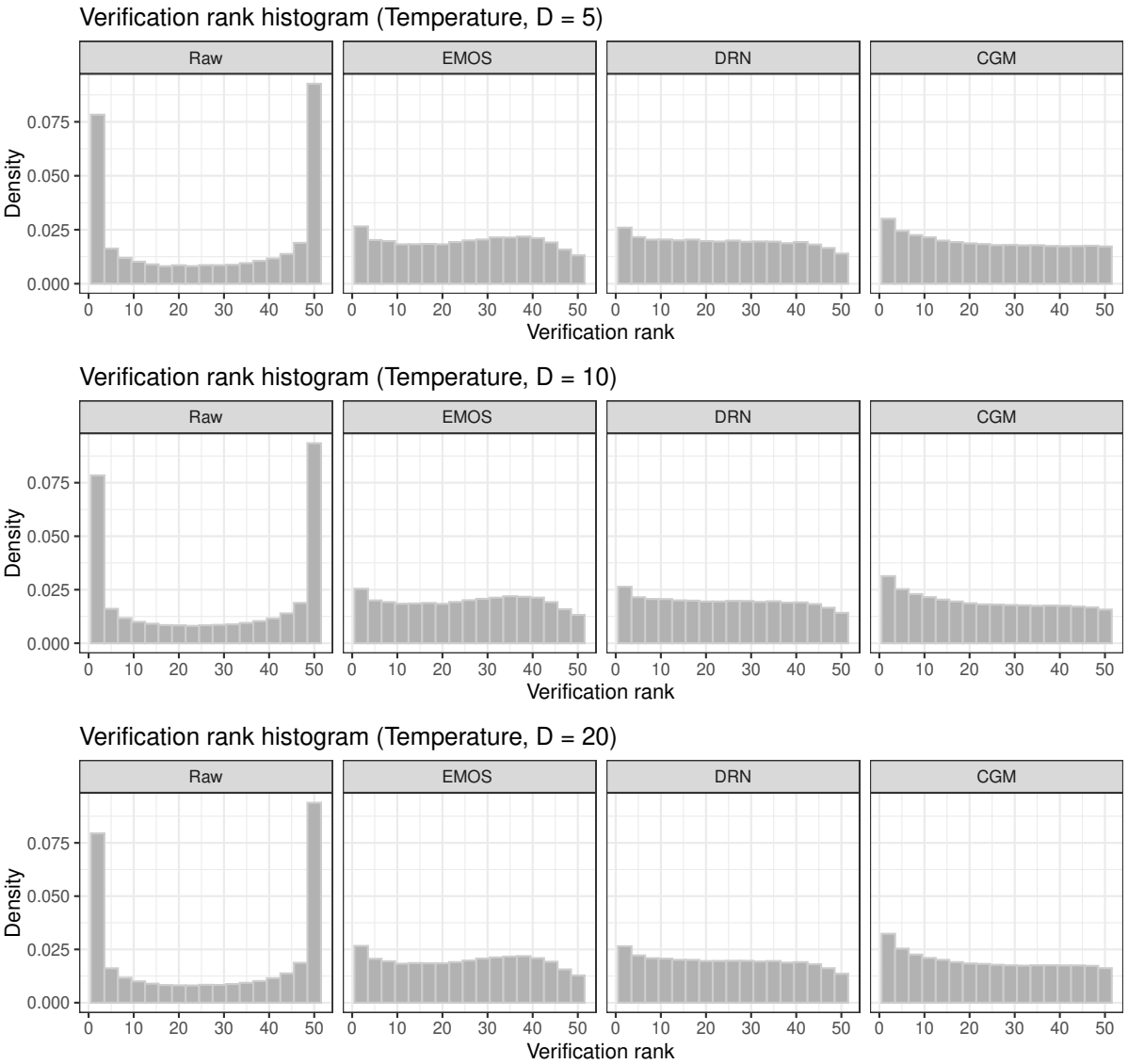

Figure 8: Verification rank histograms of different univariate post-processing methods, CGM, and the raw ensemble forecasts.

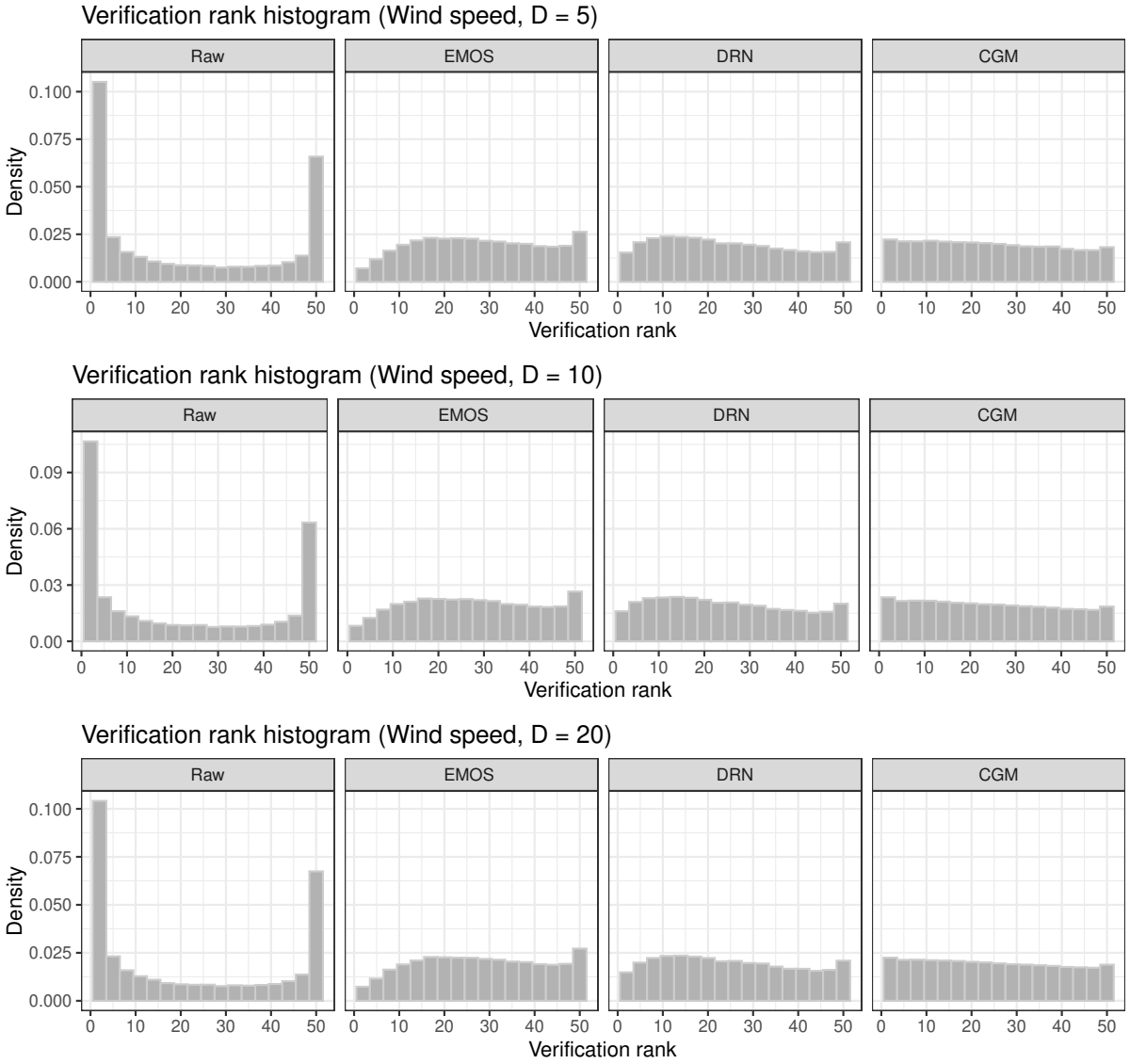

Figure 9: As Figure 8, but for wind speed.

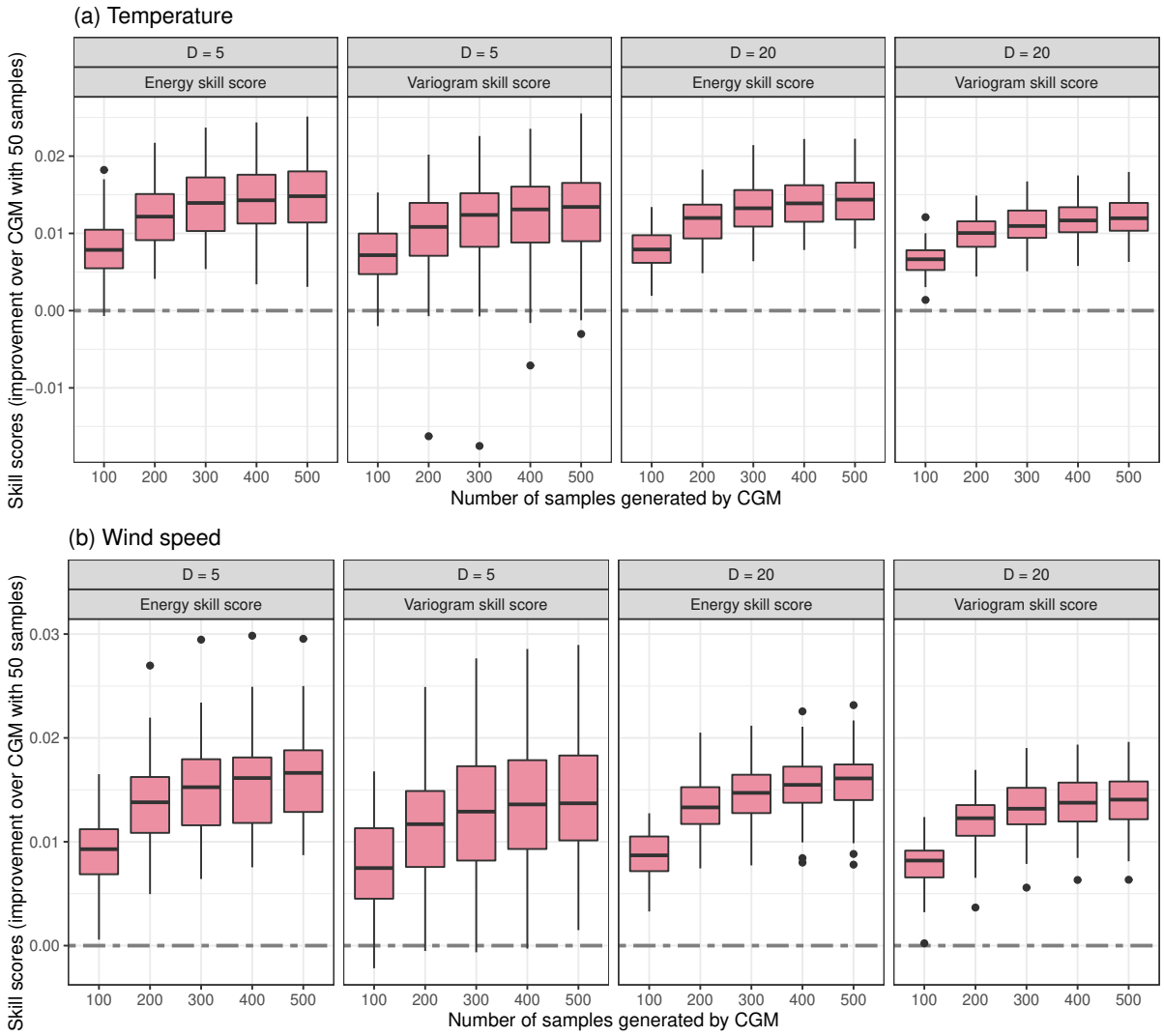

Figure 10: Boxplots of the energy skill scores and variogram skill scores of CGM forecasts with different numbers of samples generated from the multivariate post-processed forecast distribution, for (a) temperature and (b) wind speed, over the 100 repetitions of the experiment with different sets of stations. The CGM approach with $n_{\text{out}} = 50$ is used as reference forecast and we consider the cases $D = 5$ and $D = 20$.

Table 1: Proportion of pair-wise Diebold-Mariano tests for the temperature forecasts indicating statistically significant ES or VS differences after applying a Benjamini-Hochberg procedure to account for multiple testing for a nominal level of 0.05 of the corresponding one-sided tests. The (i, j)-entry in the i-th row and j-th column indicates the proportion of tests where the null hypothesis of equal predictive performance of the corresponding one-sided DM test is rejected in favor of the model in the i-th row when compared to the model in the j-th column. The remainder of the sum of (i, j)- and (j, i)-entry to 1 is the proportion of tests where the score differences are not significant. We consider the case $D = 5$ here.

Energy score					
	EMOS+ECC	EMOS+GCA	DRN+ECC	DRN+GCA	CGM
EMOS+ECC		0.02	0.00	0.00	0.00
EMOS+GCA	0.08		0.00	0.00	0.00
DRN+ECC	1.00	1.00		0.00	0.06
DRN+GCA	1.00	1.00	0.00		0.00
CGM	1.00	1.00	0.06	0.08	
Variogram score					
	EMOS+ECC	EMOS+GCA	DRN+ECC	DRN+GCA	CGM
EMOS+ECC		0.01	0.00	0.00	0.00
EMOS+GCA	0.75		0.07	0.00	0.00
DRN+ECC	0.86	0.43		0.00	0.00
DRN+GCA	0.98	0.97	0.86		0.00
CGM	0.99	0.99	0.97	0.73	

Table 2: As Table 1, but for wind speed.

Energy score					
	EMOS+ECC	EMOS+GCA	DRN+ECC	DRN+GCA	CGM
EMOS+ECC		0.00	0.00	0.00	0.00
EMOS+GCA	0.18		0.00	0.00	0.00
DRN+ECC	0.87	0.84		0.00	0.00
DRN+GCA	0.87	0.85	0.27		0.00
CGM	1.00	1.00	0.97	0.96	
Variogram score					
	EMOS+ECC	EMOS+GCA	DRN+ECC	DRN+GCA	CGM
EMOS+ECC		0.00	0.02	0.00	0.00
EMOS+GCA	0.89		0.48	0.00	0.00
DRN+ECC	0.66	0.28		0.00	0.00
DRN+GCA	0.96	0.69	0.87		0.00
CGM	1.00	0.98	1.00	0.92	

Table 3: As Table 1, but for temperature in the case $D = 20$.

Energy score					
	EMOS+ECC	EMOS+GCA	DRN+ECC	DRN+GCA	CGM
EMOS+ECC		0.00	0.00	0.00	0.00
EMOS+GCA	0.04		0.00	0.00	0.00
DRN+ECC	1.00	1.00		0.04	0.26
DRN+GCA	1.00	1.00	0.03		0.21
CGM	1.00	1.00	0.07	0.07	
Variogram score					
	EMOS+ECC	EMOS+GCA	DRN+ECC	DRN+GCA	CGM
EMOS+ECC		0.00	0.00	0.00	0.00
EMOS+GCA	0.75		0.00	0.00	0.00
DRN+ECC	1.00	0.97		0.00	0.00
DRN+GCA	1.00	1.00	0.98		0.01
CGM	0.99	0.99	0.84	0.67	

Table 4: As Table 1, but for wind speed in the case $D = 20$.

Energy score					
	EMOS+ECC	EMOS+GCA	DRN+ECC	DRN+GCA	CGM
EMOS+ECC		0.00	0.00	0.00	0.00
EMOS+GCA	0.15		0.00	0.00	0.00
DRN+ECC	1.00	1.00		0.00	0.00
DRN+GCA	1.00	1.00	0.18		0.00
CGM	1.00	1.00	1.00	1.00	
Variogram score					
	EMOS+ECC	EMOS+GCA	DRN+ECC	DRN+GCA	CGM
EMOS+ECC		0.00	0.00	0.00	0.00
EMOS+GCA	0.95		0.25	0.00	0.00
DRN+ECC	0.98	0.63		0.00	0.00
DRN+GCA	1.00	0.99	0.96		0.00
CGM	1.00	1.00	1.00	1.00	

Table 5: As Table 1, but comparing CGM forecasts with different numbers of samples generated from the multivariate post-processed forecast distribution. Here we consider the temperature forecasts in the case $D = 5$.

Energy score

#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.62		0.00	0.00	0.00	0.00
200	0.83	0.26		0.00	0.00	0.00
300	0.90	0.52	0.02		0.00	0.00
400	0.92	0.61	0.06	0.02		0.00
500	0.92	0.60	0.13	0.05	0.00	

Variogram score

#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.36		0.00	0.00	0.00	0.00
200	0.56	0.04		0.00	0.00	0.00
300	0.70	0.05	0.00		0.00	0.00
400	0.65	0.19	0.00	0.00		0.00
500	0.74	0.22	0.00	0.00	0.00	

Table 6: As Table 5, but for wind speed.

Energy score

#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.88		0.00	0.00	0.00	0.00
200	0.97	0.61		0.00	0.00	0.00
300	0.99	0.79	0.04		0.00	0.00
400	1.00	0.82	0.23	0.00		0.00
500	1.00	0.87	0.26	0.00	0.04	

Variogram score

#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.42		0.00	0.00	0.00	0.00
200	0.67	0.23		0.00	0.00	0.00
300	0.72	0.22	0.00		0.00	0.00
400	0.72	0.27	0.03	0.00		0.00
500	0.77	0.28	0.04	0.03	0.00	

Table 7: As Table 5, but for temperature in the case $D = 10$.

Energy score

#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.81		0.00	0.00	0.00	0.00
200	0.95	0.53		0.00	0.00	0.00
300	0.97	0.60	0.04		0.00	0.00
400	0.98	0.61	0.16	0.00		0.00
500	0.98	0.70	0.13	0.00	0.00	

Variogram score

#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.74		0.00	0.00	0.00	0.00
200	0.90	0.47		0.00	0.00	0.00
300	0.92	0.60	0.05		0.00	0.00
400	0.91	0.63	0.02	0.00		0.00
500	0.90	0.71	0.00	0.00	0.00	

Table 8: As Table 5, but for wind speed in the case $D = 10$.

Energy score

#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.98		0.00	0.00	0.00	0.00
200	1.00	0.83		0.00	0.00	0.00
300	1.00	0.94	0.32		0.00	0.00
400	1.00	0.95	0.59	0.08		0.00
500	1.00	0.96	0.60	0.17	0.01	

Variogram score

#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.90		0.00	0.00	0.00	0.00
200	0.94	0.66		0.00	0.00	0.00
300	0.98	0.77	0.14		0.00	0.00
400	0.99	0.83	0.31	0.05		0.00
500	0.99	0.87	0.38	0.08	0.00	

Table 9: As Table 5, but for temperature in the case $D = 20$.

Energy score						
#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.87		0.00	0.00	0.00	0.00
200	0.99	0.64		0.00	0.00	0.00
300	1.00	0.79	0.12		0.00	0.00
400	1.00	0.86	0.23	0.01		0.00
500	1.00	0.92	0.39	0.04	0.00	

Variogram score						
#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.98		0.00	0.00	0.00	0.00
200	1.00	0.70		0.00	0.00	0.00
300	1.00	0.88	0.15		0.00	0.00
400	1.00	0.89	0.40	0.00		0.00
500	1.00	0.90	0.54	0.08	0.01	

Table 10: As Table 5, but for wind speed in the case $D = 20$.

Energy score						
#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	1.00		0.00	0.00	0.00	0.00
200	1.00	0.94		0.00	0.00	0.00
300	1.00	0.97	0.33		0.00	0.00
400	1.00	0.97	0.59	0.09		0.00
500	1.00	0.98	0.69	0.22	0.03	

Variogram score						
#	50	100	200	300	400	500
50		0.00	0.00	0.00	0.00	0.00
100	0.97		0.00	0.00	0.00	0.00
200	0.99	0.92		0.00	0.00	0.00
300	1.00	0.93	0.36		0.00	0.00
400	1.00	0.96	0.57	0.07		0.00
500	1.00	0.99	0.69	0.32	0.02	

cases.

2.2.2 Copula scores

As noted in the main paper, a major challenge in the evaluation of multivariate forecasts is the distinction between contributions of improvements in the marginal distributions and the multivariate dependencies to the overall forecast quality. The univariate results presented above and in the main paper indicate that the DRN-based multivariate post-processing methods perform notably better in the calibration of marginal distributions than the EMOS-based methods, which is reflected on the improvement of multivariate scores given that they employ the same copula reordering (ECC or GCA) approach. For CGM forecasts, it is not trivial to assess to what extent the dependence structure improves the scores when compared with other methods.

Ziel and Berk (2019) propose the copula energy score and the copula variogram score as a novel approach to address this challenge. The copula scores are a new class of multivariate proper scoring rules that focus on the dependency structure of the multivariate forecast distribution. For more exact definitions, as well as further mathematical details and illustrations we refer to Ziel and Berk (2019).

We apply the copula energy score and the copula variogram score to assess the forecast performance of different multivariate post-processing methods, and the corresponding skill scores taking EMOS+ECC as reference are shown in Figures 11 and 12. In general, the performance of copula-based methods depends on the corresponding copulas, with GCA notably outperforming ECC. The choice of the univariate post-processing method only has a minor effect on the performance of copula-based methods (as expected), with DRN perhaps surprisingly leading to slightly worse results than EMOS. CGM shows comparable or slightly better performance than the GCA-based approaches, which suggests that the dependence structures of the CGM forecasts better match those in the observations, when compared with the best-performing GCA-based approaches.

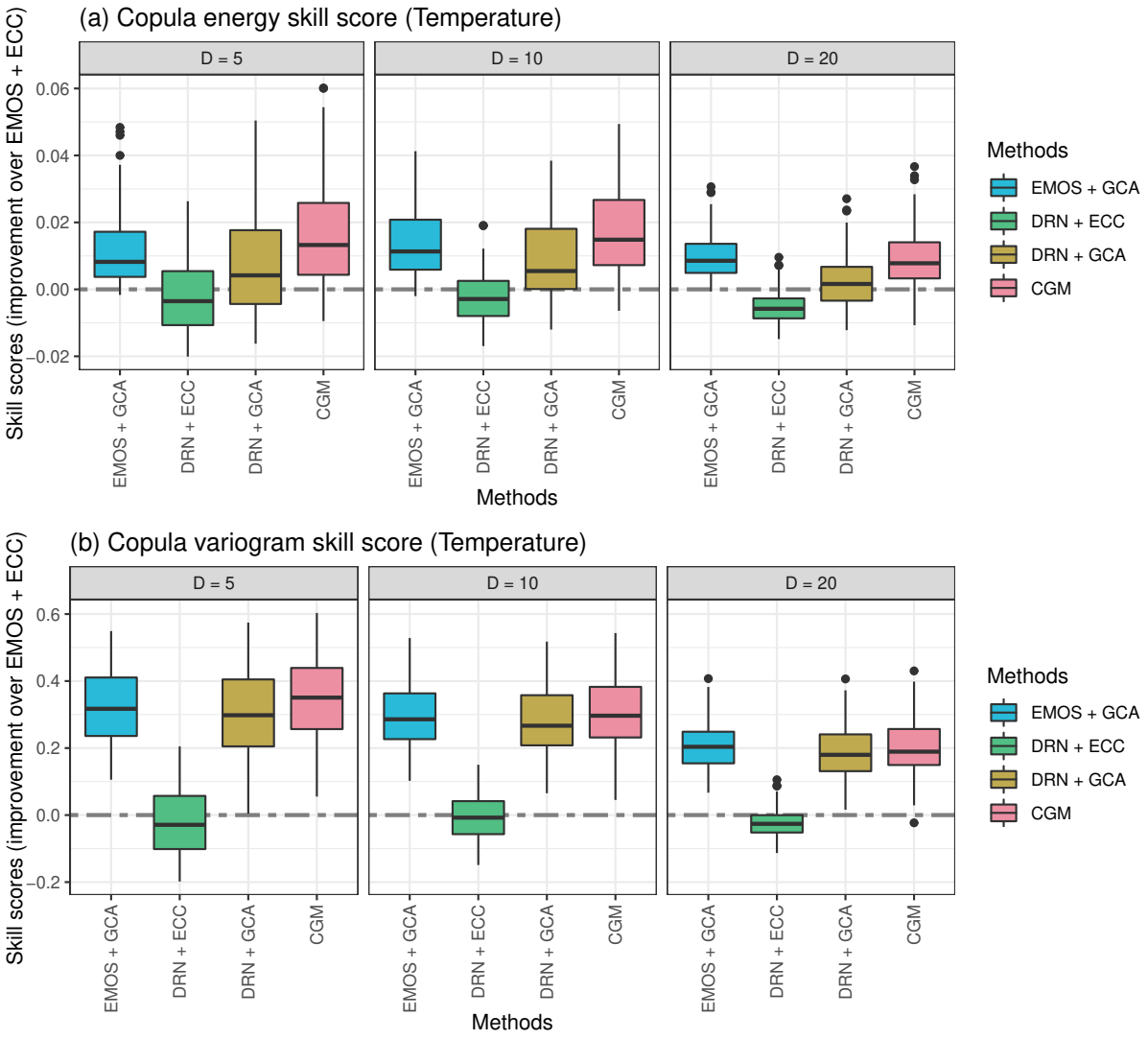

Figure 11: Boxplots of (a) copula energy skill scores and (b) copula variogram skill scores of different multivariate post-processing methods for temperature across the 100 repetitions of the experiment with different sets of stations. EMOS+ECC is used as reference forecast in both cases.

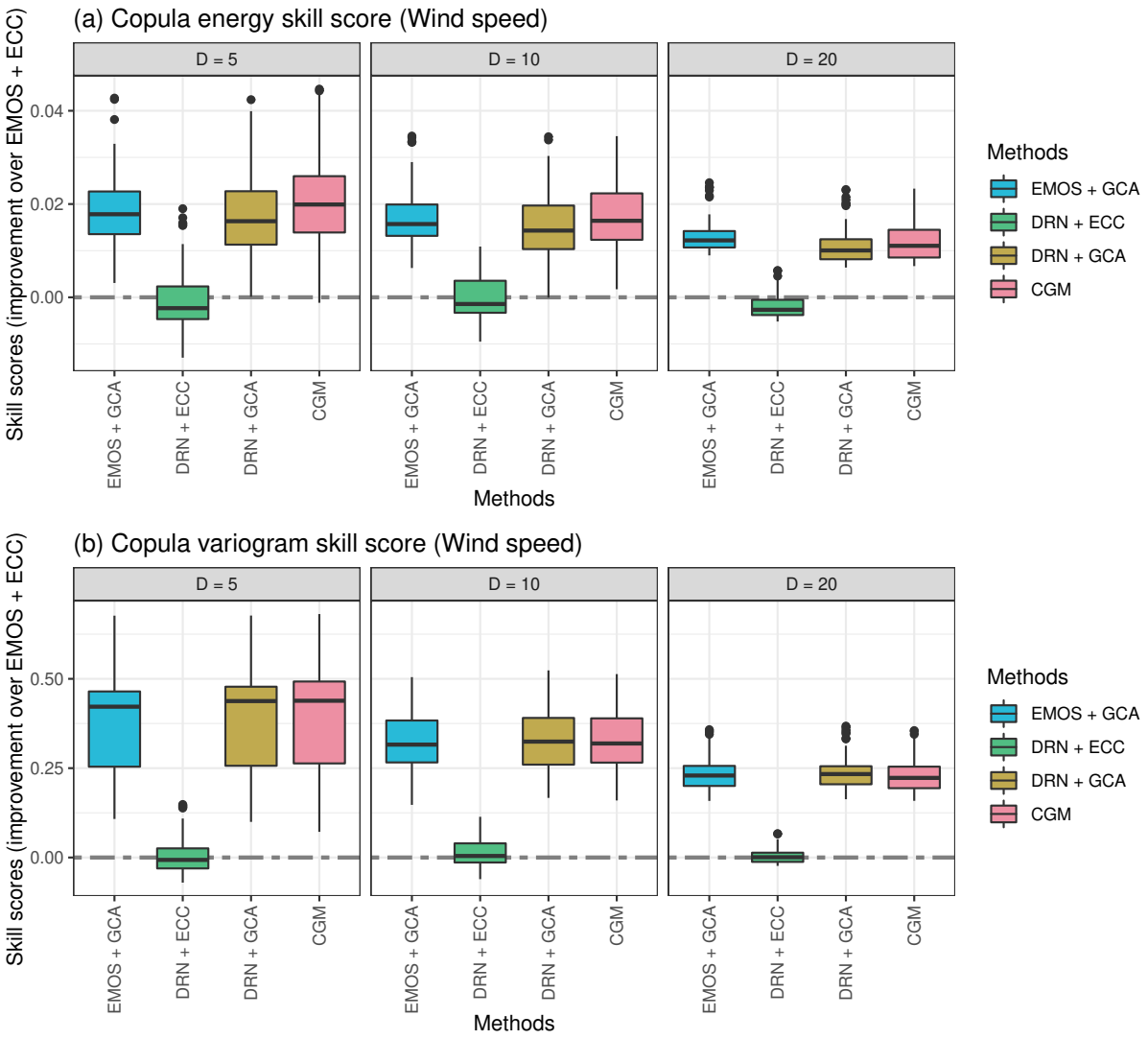

Figure 12: As Figure 11, but for wind speed.

2.2.3 Weighted multivariate scores

As discussed in the main paper, one benefit from the CGM approach is the option to efficiently generate a large number of samples from the post-processed multivariate distribution. This should in principle prove advantageous in correctly representing multivariate extreme events. To assess the multivariate performance for predicting extreme events we apply the weighted multivariate scoring rules proposed by Allen et al. (2022). Specifically, we consider the threshold-weighted energy score and variogram score in both their localizing ("twES-loc", "twVS-loc") and non-localizing ("twES", "twVS") variants, as well as the vertically re-scaled energy score and variogram score ("vrES", "vrVS"). For the exact definitions, and detailed mathematical illustration and simulation studies we refer to Allen et al. (2022). As a threshold, we here select the 95%-quantile of the observed values to determine extreme events. This corresponds to temperatures exceeding 17.5°C and wind speeds exceeding 8.0 m/s. Note that we here choose fixed thresholds to gain an overall perspective of the multivariate performance. In a more detailed future study, it might be interesting to account for seasonally adjusted and location-specific climatologies when determining the threshold values.

The mean scores over the 100 experiments are shown in Tables 11 and 12. We here additionally include the results for the CGM forecast with 500 samples for comparison. While the results show some variations across the two variables and the different weighted scores, the CGM variant with 500 samples generally provide the best forecasts. By contrast, the default CGM variant with 50 samples usually performs very similar to the best of the two-step methods (DRN+GCA) for temperature, and slightly better than DRN+GCA for wind speed.

2.2.4 Multivariate rank histograms

Thorarinsdottir et al. (2016) propose several extensions of univariate rank histograms towards multivariate versions that allow for assessing multivariate calibration. We here consider the multivariate rank, average rank and band depth rank histograms, and refer to Thorarinsdottir et al. (2016) for the exact definitions and details of the interpretation of different histogram shapes.

Figures 13, 14, 15, 16, 17 and 18 show the corresponding histograms. Overall, the CGM forecasts and the GCA-based forecasts show superior multivariate calibration performance to ECC-based forecasts, but neither of them consistently achieves uniformly distributed histogram across all types of ranks. In terms of the band depth rank, the ECC-based approaches lead to an over-estimation of the correlation among the ensemble members for both temperature and wind speed forecasts, which can also be observed in the average rank histograms. In general, there is no multivariate post-processing method that performs well consistently over all types of ranks, indicating that rankings of the different models will strongly depend on the employed notion of multivariate calibration.

References

Allen, S., Ginsbourger, D. and Ziegel, J. (2022). Evaluating forecasts for high-impact events using transformed kernel scores. Preprint, URL `https://arxiv.org/abs/2202.12732`.

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

Ziel, F. and Berk, K. (2019). Multivariate forecasting evaluation: On sensitive and strictly proper scoring rules. Preprint, URL `https://arxiv.org/abs/1910.07325`.

Table 11: Mean weighted multivariate scores of different post-processing methods for temperature, averaged over the 100 repetitions of the simulation experiment. The weighted versions of the energy score (ES) and the variogram score (VS) are considered. We scale the scores by 10 where indicated by "×10" for better interpretation. The best scores are highlighted in bold.

Score	Raw ens.	EMOS+ ECC	EMOS+ GCA	DRN+ ECC	DRN+ GCA	CGM	CGM (500 samples)
				$D = 5$			
ES	2.81	2.27	2.27	1.97	1.97	1.97	**1.94**
twES	0.172	0.143	0.143	0.130	0.130	0.129	**0.127**
twES-loc (×10)	1.18	0.784	0.756	0.726	0.698	0.704	**0.690**
vrES	1.04	0.552	0.522	0.502	0.473	0.477	**0.468**
VS	8.22	4.81	4.36	4.12	3.74	3.50	**3.46**
twVS	0.876	0.699	0.644	0.590	0.566	0.561	**0.550**
vrVS	0.367	0.308	0.300	0.295	0.282	0.283	**0.277**
				$D = 10$			
ES	4.22	3.37	3.37	2.91	2.90	2.91	**2.87**
twES	0.254	0.210	0.208	0.187	0.187	0.185	**0.182**
twES-loc (×10)	1.44	0.715	0.671	0.654	**0.601**	0.626	0.617
vrES	1.25	0.434	0.398	0.400	**0.353**	0.366	0.360
VS	39.0	22.6	21.0	19.5	18.0	16.9	**16.7**
twVS	3.95	3.11	2.90	2.61	2.54	2.45	**2.41**
vrVS	1.36	0.865	0.820	0.815	**0.752**	0.767	0.756
				$D = 20$			
ES	6.09	4.87	4.87	4.21	4.22	4.26	**4.20**
twES	0.367	0.300	0.298	0.272	0.272	0.272	**0.267**
twES-loc (×10)	1.85	0.640	0.592	0.584	**0.527**	0.570	0.563
vrES	1.56	0.384	0.346	0.339	**0.292**	0.321	0.317
VS	153	96.7	92.8	85.0	80.7	77.8	**76.9**
twVS	15.8	12.6	12.1	10.8	10.7	10.5	**10.4**
vrVS	5.06	2.29	2.19	2.15	**1.97**	2.09	2.07

Table 12: As Figure 11, but for wind speed.

Score	Raw ens.	EMOS+ ECC	EMOS+ GCA	DRN+ ECC	DRN+ GCA	CGM	CGM (500 samples)
			$D = 5$				
ES	2.44	1.69	1.68	1.56	1.55	1.44	**1.42**
twES	0.230	0.168	0.166	0.149	0.149	0.133	**0.131**
twES-loc (×10)	0.460	0.337	0.337	0.330	0.309	0.304	**0.301**
vrES (×10)	1.40	0.955	0.966	0.941	0.883	0.860	**0.853**
VS	9.49	4.37	4.00	4.01	3.66	3.31	**3.26**
twVS	1.5	0.978	0.960	0.810	0.800	0.687	**0.675**
vrVS	0.221	0.178	0.179	0.167	0.160	0.160	**0.157**
			$D = 10$				
ES	3.67	2.55	2.53	2.31	2.30	2.16	**2.12**
twES	0.504	0.369	0.365	0.323	0.323	0.288	**0.284**
twES-loc (×10)	0.541	0.325	0.334	0.311	**0.276**	0.314	0.313
vrES (×10)	1.54	0.759	0.775	0.771	**0.648**	0.754	0.757
VS	39.7	20.2	19.0	18.0	16.9	15.4	**15.2**
twVS	9.11	5.87	5.79	4.75	4.73	3.97	**3.92**
vrVS	0.866	0.506	0.518	0.491	**0.422**	0.514	0.517
			$D = 20$				
ES	5.04	3.52	3.51	3.23	3.22	3.04	**2.99**
twES	0.587	0.435	0.428	0.386	0.387	0.352	**0.346**
twES-loc (×10)	0.266	0.118	0.121	0.124	0.110	0.101	**0.0959**
vrES (×10)	0.865	0.271	0.275	0.290	0.249	0.228	**0.215**
VS	153	82.5	78.9	75.6	72.3	67.0	**66.0**
twVS	28.3	18.8	18.6	15.5	15.4	13.3	**13.1**
vrVS	1.34	0.612	0.628	0.619	0.557	0.511	**0.480**

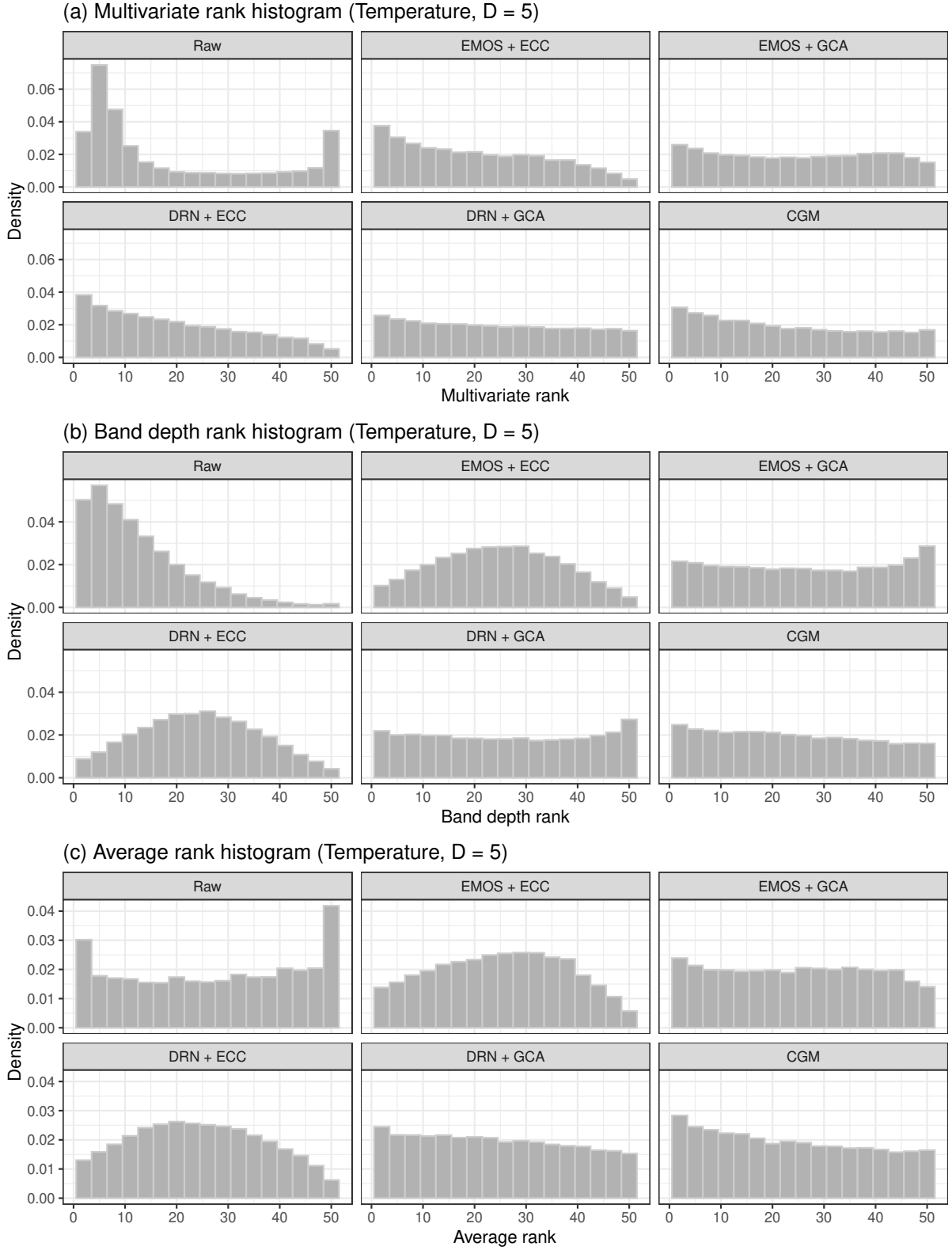

Figure 13: Histograms of (a) the multivariate rank, (b) the band depth rank, and (c) the average rank of different multivariate post-processing methods and the raw ensemble forecasts for temperature, across the 100 repetitions of the simulation experiment and the 366 days in the test set (calendar year 2016). We consider the case $D = 5$ here.

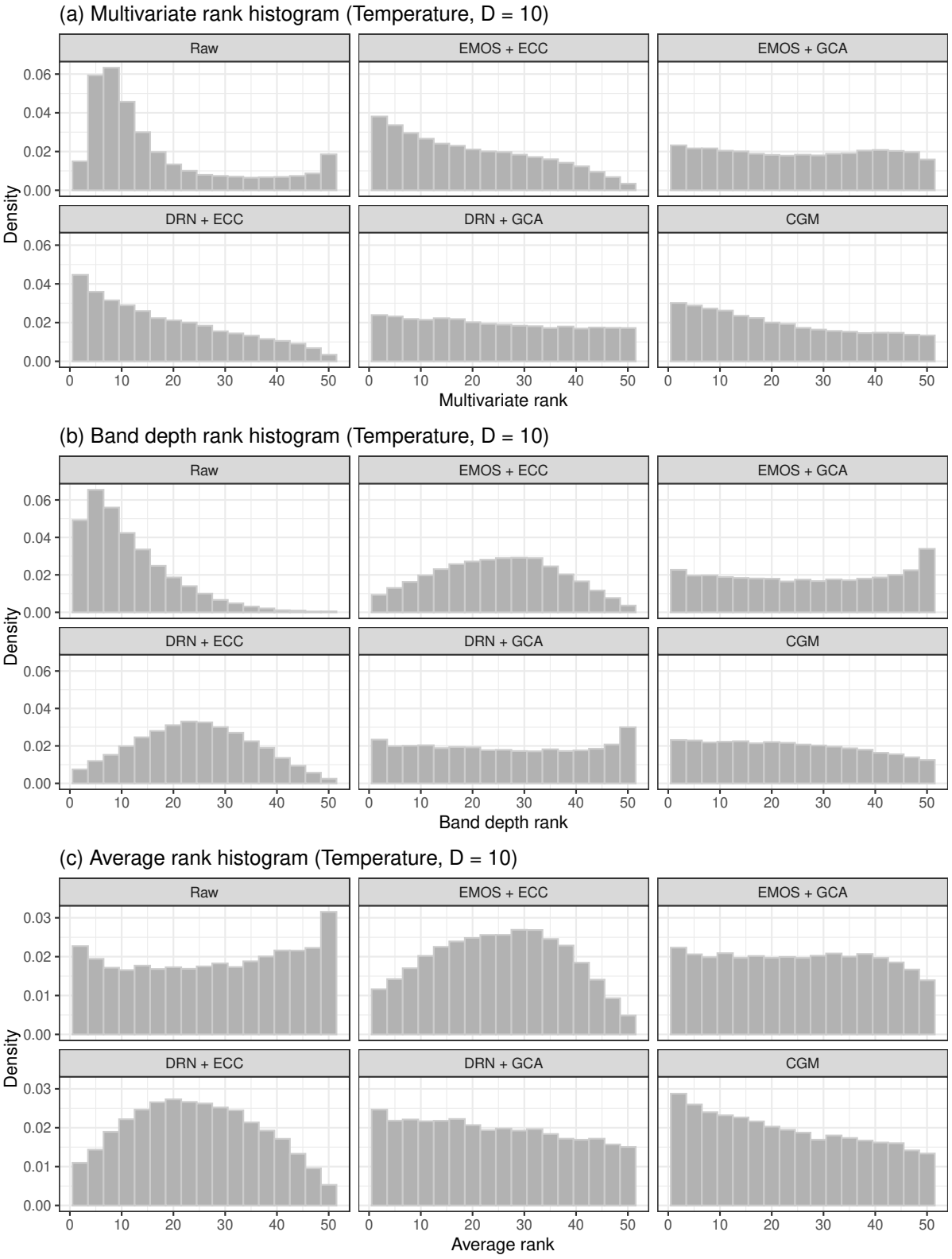

Figure 14: As Figure 13, but for temperature and $D = 10$.

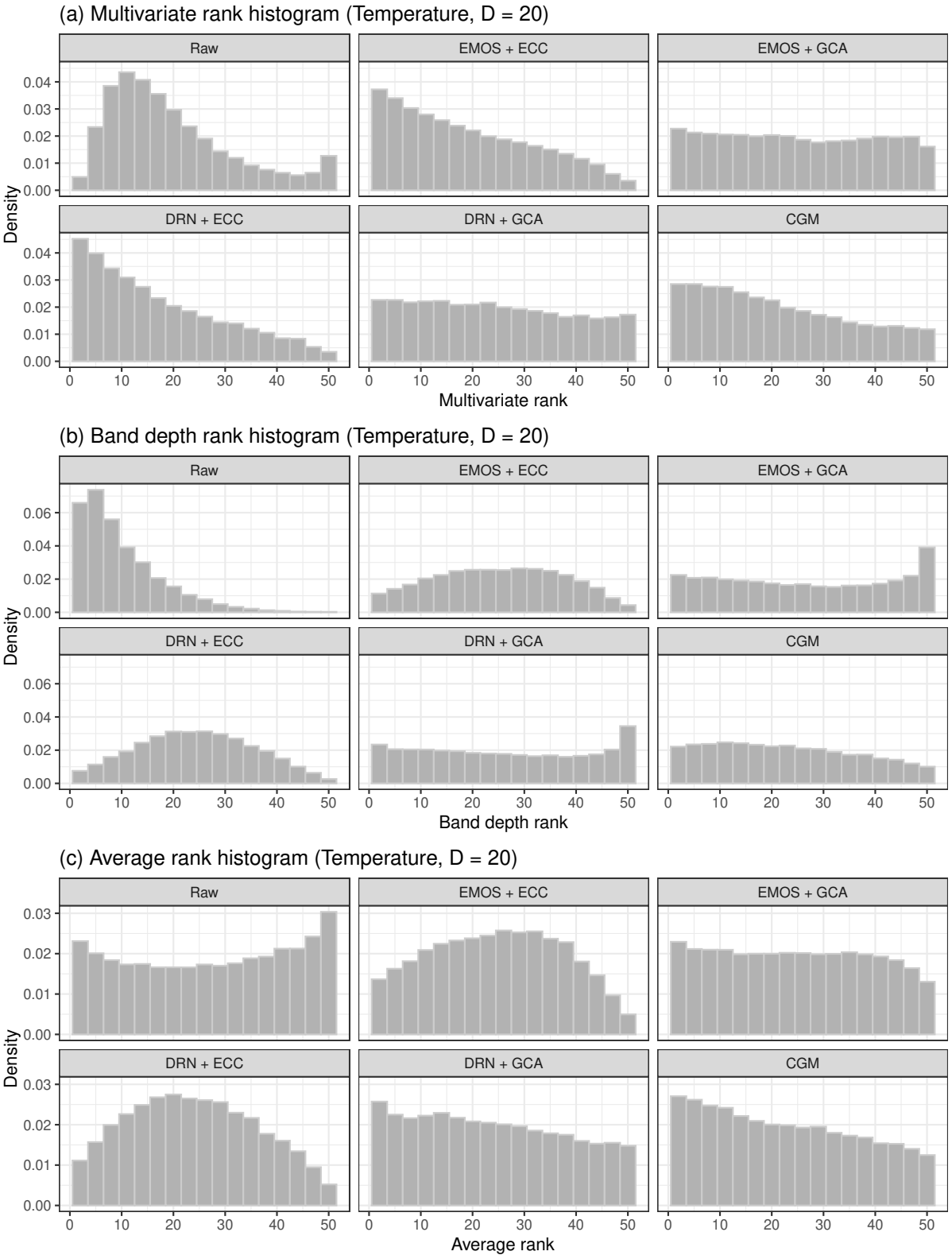

Figure 15: As Figure 13, but for temperature and $D = 20$.

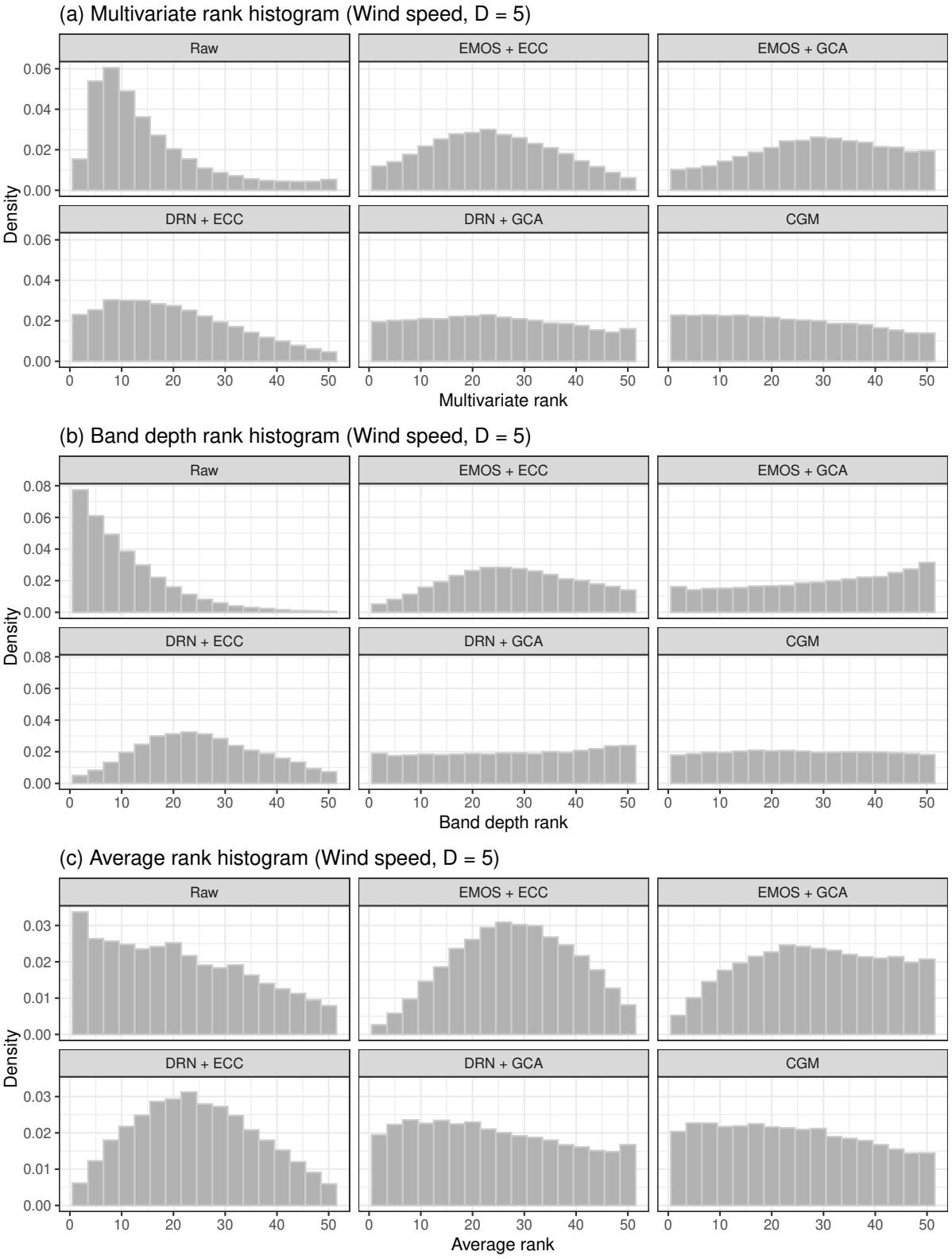

Figure 16: As Figure 13, but for wind speed.

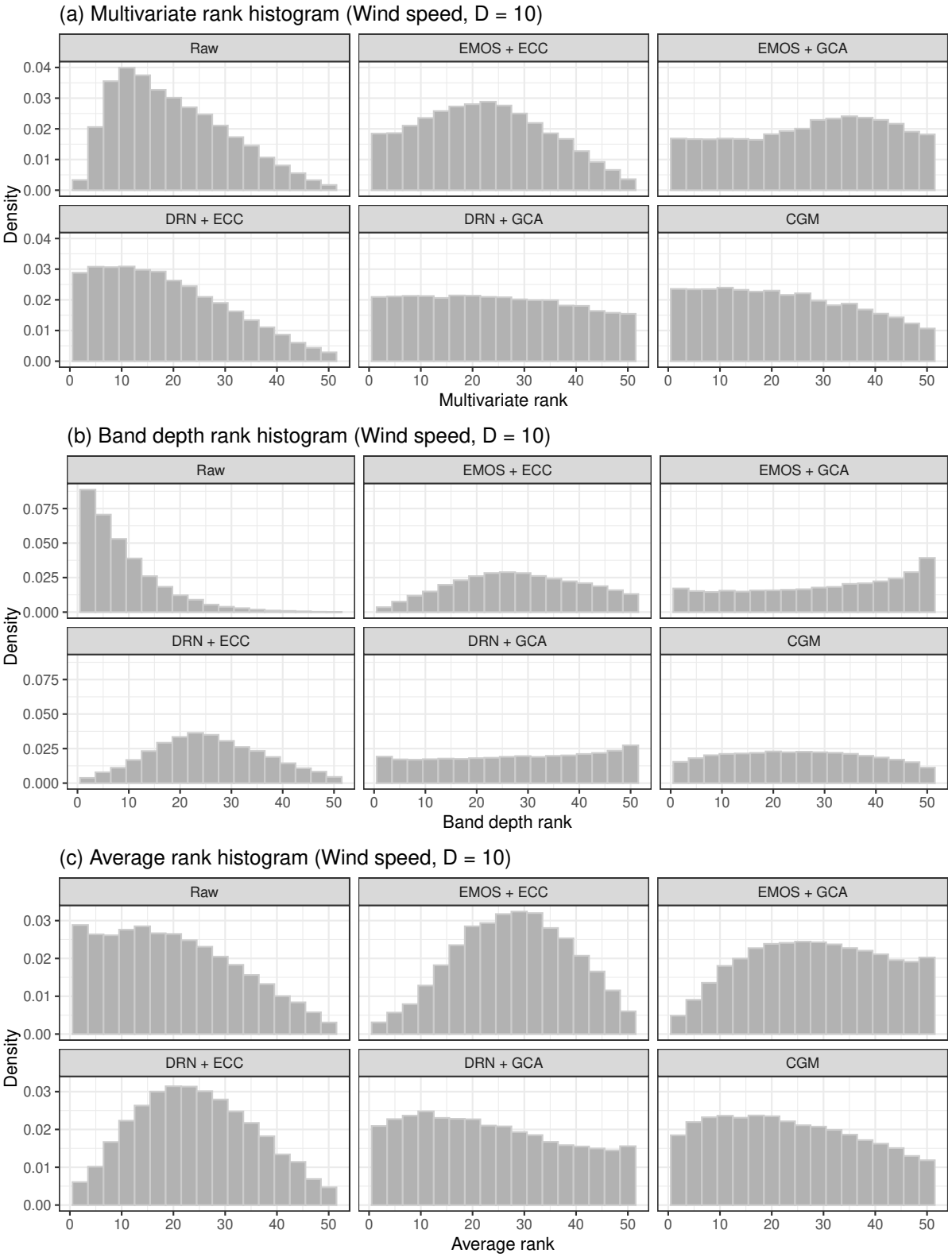

Figure 17: As Figure 13, but for wind speed and $D = 10$.

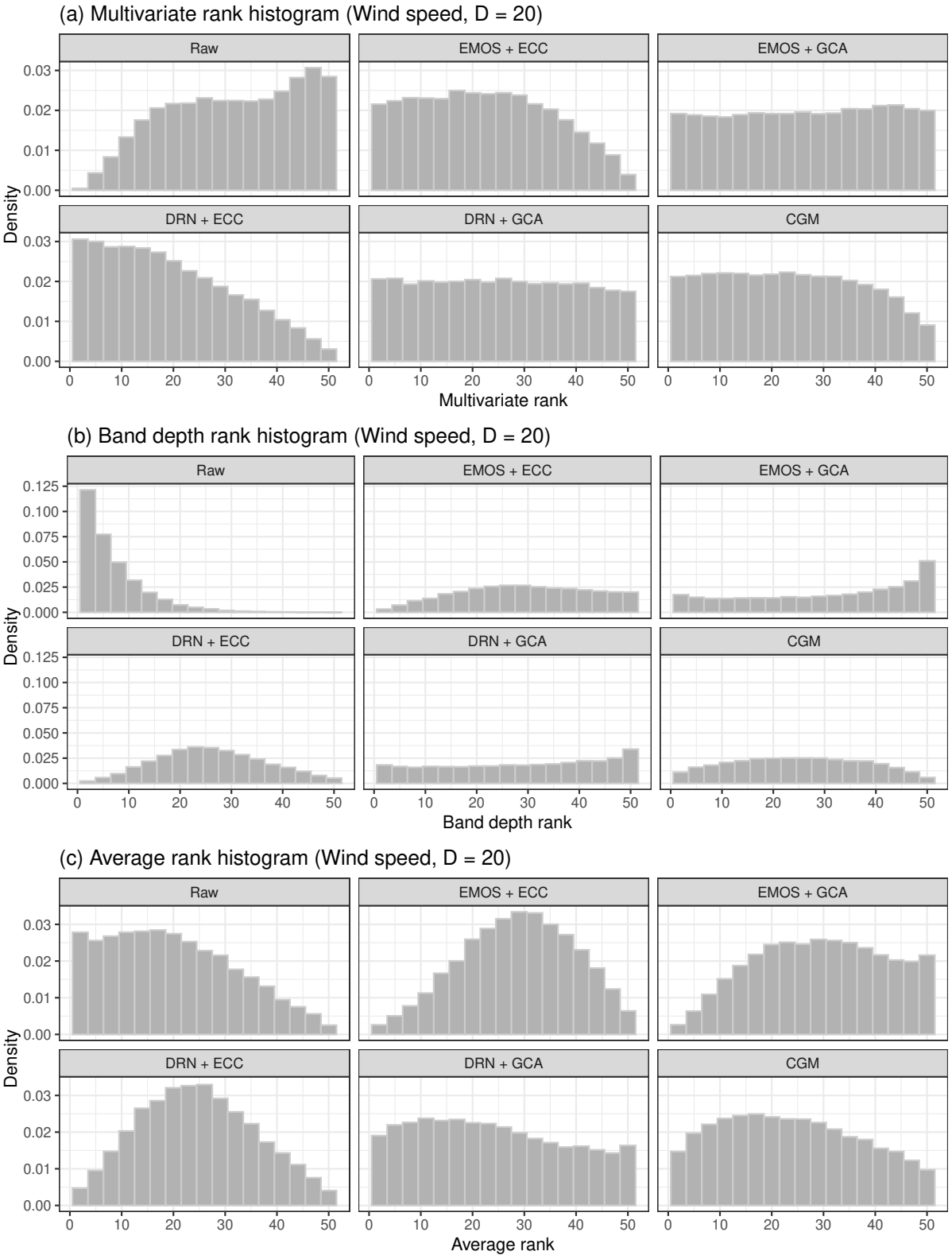

Figure 18: As Figure 13, but for wind speed and $D = 20$.