CORRIGENDUM

BENEDIKT SCHULZ[a] AND SEBASTIAN LERCH[a,b]

[a] *Karlsruhe Institute of Technology, Karlsruhe, Germany*
[b] *Heidelberg Institute for Theoretical Studies, Heidelberg, Germany*

(Manuscript received 16 January 2023, in final form 22 March 2023, accepted 23 March 2023)

In Schulz and Lerch (2022), there was an error in the calculation of the Diebold–Mariano (DM; Diebold and Mariano 1995) test statistics that affected Table 5 and Fig. 7 of that paper but not the overall conclusions drawn in the study. In section d of appendix A, the third equation (the variance term) is incorrect because a square root transformation in the variance term is missing, and it instead should read as

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}[S(F_i, y_i) - S(G_i, y_i)]^2$$

(i.e., a square or power of 2 has been added to the left-hand-side term). Note that we estimate the variance without centering [following, e.g., Gneiting and Katzfuss (2014)]. With centering, the DM test is equivalent to a standard t test for the score difference being equal to 0 in expectation. Both variants are valid estimators of the variance under the null hypothesis. We refer to Jordan (2016) for details.[1]

The code was also subject to this error,[2] and hence the original Table 5 and Fig. 7 are recalculated using the correct formula. Corrected versions are displayed here in Table 5 and Fig. 7. The conclusions drawn in Schulz and Lerch (2022) still hold, but the numbers stated in one paragraph need to be updated. The third paragraph of section 4d should now read as follows:

> We find that the observed score differences are statistically significant for a high ratio of stations and lead times; see Table 5. In particular, DRN and BQN significantly outperform the basic models at more than 94%, and even significantly outperform QRF and EMOS-GB at more than 50% of all combinations of stations and lead times. Among the locally estimated methods, QRF performs best but only provides significant improvements over the NN-based methods for around 1% of the cases.

That is, the three percentages have been reduced from 97%, 80%, and 5% to 94%, 50%, and 1%, respectively.

In general, the ratio of significant tests in Table 5 decreases. Especially the comparison of the network methods DRN, BQN, and HEN are affected, where DRN and BQN now perform significantly better than the other method at less than 2% instead of at ~44%–46%. For HEN, QRF, and EMOS-GB, the ratio of significant tests for the superiority of DRN and BQN also decreases by ~25–40 percentage points.

In comparing the original Fig. 7 in Schulz and Lerch (2022) with the revised Fig. 7, it is seen that the size of symbols decreases because of the decreased ratio of significant tests. Note that the best-performing method at each location is not affected and does not change.

[1] We thank Ron McTaggart-Cowan, Michael Scheuerer, and Alexander Jordan for constructive discussions on the different variants of the DM test.

[2] In addition to this *Corrigendum*, we updated the code on the corresponding GitHub repository (https://github.com/benediktschulz/paper_pp_wind_gusts).

Corresponding author: Benedikt Schulz, benedikt.schulz2@kit.edu

DOI: 10.1175/MWR-D-23-0010.1

TABLE 5. Ratio of lead time–station combinations (%) where pairwise DM tests indicate statistically significant CRPS differences after applying a Benjamini–Hochberg procedure to account for multiple testing for a nominal level of $\alpha = 0.05$ of the corresponding one-sided tests. The (i, j) entry in the ith row and jth column indicates the ratio of cases where the null hypothesis of equal predictive performance of the corresponding one-sided DM test is rejected in favor of the model in the ith row when compared with the model in the jth column. The remainder of the sum of (i, j) and (j, i) entry to 100% is the ratio of cases for which the score differences are not significant.

	EPC	EPS	EMOS	MBM	IDR	EMOS-GB	QRF	DRN	BQN	HEN
EPC		5.4	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
EPS	78.9		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EMOS	99.3	99.9		84.8	51.1	0.0	0.0	0.0	0.0	0.0
MBM	99.3	99.8	0.0		5.7	0.0	0.0	0.0	0.0	0.0
IDR	98.7	99.2	0.0	1.7		0.0	0.0	0.0	0.0	0.0
EMOS-GB	100.0	99.9	69.5	87.5	87.3		0.5	0.2	0.2	1.2
QRF	100.0	99.9	70.3	88.0	91.9	6.1		1.0	1.1	2.7
DRN	99.9	100.0	94.2	97.7	97.3	58.0	52.8		1.8	44.7
BQN	99.9	100.0	94.2	97.3	97.4	56.6	53.1	1.0		43.4
HEN	99.6	99.9	87.0	94.2	93.6	29.6	26.1	0.1	0.0	

REFERENCES

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263, https://doi.org/10.1080/07350015.1995.10524599.

Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831.

Jordan, A., 2016: Facets of forecast evaluation. Ph.D. dissertation, Karlsruhe Institute of Technology, 112 pp., https://doi.org/10.5445/IR/1000063629.

Schulz, B., and S. Lerch, 2022: Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Wea. Rev.*, **150**, 235–257, https://doi.org/10.1175/MWR-D-21-0150.1.

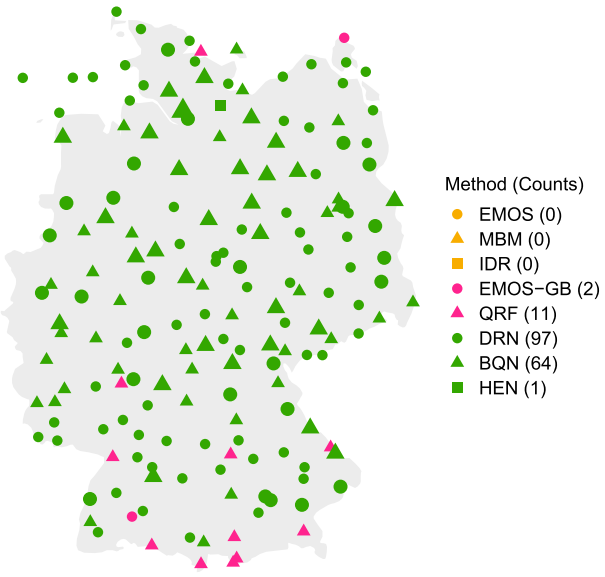

FIG. 7. Best method at each station in terms of the CRPS, averaged over all lead times. The point sizes indicate the level of statistical significance of the observed CRPS differences relative to the methods only from the other groups of methods for all lead times. Three different point sizes are possible, with the smallest size indicating statistically significant differences for at most 90% of the performed tests, the middle size is for up to 99%, and the largest is to 100%, meaning all differences are statistically significant.