



Post-processing ensemble weather forecasts for discrete and continuous quantities in univariate and multivariate settings

Thesis for the Degree of Doctor of Philosophy (PhD)

Mária Nagy-Lakatos

Supervisor: Prof. Dr. Sándor Baran

UNIVERSITY OF DEBRECEN

Doctoral Council for Natural Sciences and Engineering

Doctoral School of Informatics

Debrecen, 2025

Hereby I declare that I prepared this thesis within the Doctoral Council for Natural Sciences and Engineering, Doctoral School of Informatics, University of Debrecen in order to obtain a PhD Degree in Engineering at Debrecen University.

The results published in the thesis are not reported in any other PhD theses.

Debrecen, 202_____

signature of the candidate

Hereby I confirm that Mária Nagy-Lakatos candidate conducted her studies with my supervision within the Theoretical Foundation and Applications of Information Technology and Stochastic Systems Doctoral Program of the Doctoral School of Informatics between 2021 and 2025. The independent studies and research work of the candidate significantly contributed to the results published in the thesis.

I also declare that the results published in the thesis are not reported in any other theses.

I support the acceptance of the thesis.

Debrecen, 202_____

signature of the supervisor

**Post-processing ensemble weather forecasts for
discrete and continuous quantities in univariate
and multivariate settings**

Dissertation submitted in partial fulfilment of the requirements for the
doctoral (PhD) degree in Informatics

Written by Mária Nagy-Lakatos, certified computer scientist

Prepared in the framework of the Doctoral School of Informatics,
University of Debrecen
(Theoretical Foundation and Applications of Information Technology and
Stochastic Systems programme)

Dissertation supervisor: Prof. Dr. Sándor Baran

The official opponents of the dissertation:

Dr.
Dr.

The evaluation board:

chairperson: Dr.
members: Dr.
Dr.
Dr.
Dr.

The date and venue of the dissertation defence: 202

Acronyms

BMA:	Bayesian Model Averaging
BS:	Brier Score
CAMS:	Copernicus Atmosphere Monitoring Service
CNTL:	Control Forecast
CRPS:	Continuous Ranked Probability Score
dECC:	Dual Ensemble Copula Coupling
DM test:	Diebold-Mariano test
ECC:	Ensemble Copula Coupling
EMOS:	Ensemble Model Output Statistics
ENS:	Ensemble Forecast
EPS:	Ensemble Prediction System
ES:	Energy Score
HRES:	High-Resolution Forecast
MAE:	Mean Absolute Error
mdSSh:	Minimum Divergence Schaake Shuffle
NWP:	Numerical Weather Prediction
PIT:	Probability Integral Transform
POLR:	Proportional Odds Logistic Regression
RI:	Reliability Index

RMSE: Root Mean Squared Error

SSh: Schaake Shuffle

simSSh: Similarity-based Schaake Shuffle

SYNOP: Surface Synoptic Observation

T2M: 2-Meter Temperature

twCRPS: Threshold-Weighted Continuous Ranked Probability Score

VS: Variogram Score

Contents

Introduction	1
1 Ensemble forecasting and post-processing	5
1.1 Ensemble forecasts	5
1.2 Ensemble model output statistics	7
1.2.1 Normal EMOS model	9
1.2.2 Truncated normal EMOS model	10
1.2.3 Censored generalized extreme value EMOS model	10
1.3 Discrete post-processing	11
1.3.1 Proportional odds logistic regression	11
1.3.2 Multilayer perceptron	11
2 Multivariate post-processing	13
2.1 Ensemble Copula Coupling	14
2.2 Dual Ensemble Copula Coupling	14
2.3 Schaake Shuffle	15
2.4 Minimum Divergence Schaake Shuffle	15
2.5 Similarity-based Schaake Shuffle	16
2.6 Assumption of independence	17
2.7 Operational implementation of multivariate post-processing	18
3 Forecast evaluation and training data	19
3.1 Verification metrics	19
3.2 Training data selection	24
4 Clustering-based spatial interpolation	27
4.1 Interpolation algorithms	28
4.2 Data	31
4.3 Results	32
4.4 Summary and conclusion	42

5	Post-processing of visibility forecasts	43
5.1	Introduction	43
5.2	Data	44
5.3	Results	46
5.3.1	Calibration of EUPPBench visibility ensemble forecasts	50
5.3.2	Post-processing ECWMF visibility ensemble forecasts using CAMS covariates	54
5.4	Conclusions	64
6	Comparison of multivariate methods	67
6.1	Data	68
6.2	Results	69
6.2.1	Univariate post-processing	69
6.2.2	Multivariate performance	71
6.3	Discussion	83
7	Summary	87
8	Összefoglalás	91
	Acknowledgements	95
	Bibliography	96
A	List of publications and conferences	109
B	ECMWF visibility ensemble calibration	113
C	Dawid-Sebastiani score	119
C.1	Definition	119
C.2	Forecast evaluation with RDSS	119

List of Figures

1.1	Calibrated normal EMOS probability density functions for the HungaroMet's 42 h temperature forecasts, for Debrecen Airport for 3 different dates, illustrating the EMOS median (red solid line), raw ensemble (green segments) and observation (blue dashed line).	8
2.1	The effect of multivariate post-processing: 24, 96, 168, and 240 h ahead ECMWF ensemble forecasts for temperature initialized at 18 November 2006; a) raw ensemble forecasts b) after individual EMOS postprocessing (losing dependence structures) c) after ECC, and d) Schaake shuffle post-processing.	17
3.1	Illustration of the continuous ranked probability score (CRPS, gold area) through three simulated forecast cases (normal CDFs) and their corresponding observations	20
4.1	Flowchart of the semi-local parameter estimation method for the observed locations and the interpolation algorithm for an unobserved site.	30
4.2	Locations of SYNOP stations and the distribution of observed and unobserved locations.	32
4.3	CRPS of post-processed and raw 10-m wind speed forecasts for observed (a) and unobserved (b) locations, and CRPSS of local and semi-local models with respect to the regional approach for observed (c) and unobserved (d) locations together with 95 % confidence intervals as functions of the lead time.	33
4.4	Proportions of observed (upper triangle) and unobserved (lower triangle) locations with significantly different mean CRPS at a 5 % level of significance for lead times 1 (a), 4 (b), 7 (c) and 10 (d) days.	34

4.5	twCRPSS values of semi-local models with respect to the regional approach for thresholds 6 m/s (a,b), 8 m/s (c,d) and 9 m/s (e,f) for observed (a,c,e) and unobserved (b,d,f) locations together with 95 % confidence intervals as functions of the lead time.	35
4.6	twCRPSS values of semi-local models with respect to the regional approach for thresholds 6 m/s (a,b), 8 m/s (c,d) and 9 m/s (e,f) for observed (a,c,e) and unobserved (b,d,f) locations together with 95 % confidence intervals as functions of the lead time.	36
4.7	Coverage (a,b) and average width (c,d) of nominal 96.23 % central prediction intervals of post-processed and raw 10-m wind speed forecasts for observed (a,c) and unobserved (b,d) locations functions of the lead time. In panels (a) and (b) the ideal coverage is indicated by the horizontal dotted line.	37
4.8	PIT histograms of post-processed and verification rank histograms of raw ensemble forecasts of 10-m wind speed for lead times 1, 4, 7 and 10 days. The ideal uniform distribution is indicated by the horizontal dotted line. The cross-hatced parts indicate the intersection of the corresponding histograms for observed and unobserved locations.	38
4.9	MAE of the median 10-m wind speed forecasts for observed (a) and unobserved (b) locations, and difference in MAE of local and semi-local models from the regional approach for observed (c) and unobserved (d) locations together with 95 % confidence intervals as functions of the lead time.	40
5.1	Locations of SYNOP observation stations corresponding to (a) EUPPBench benchmark dataset; (b) ECMWF forecasts for 2020-2021.	45
5.2	Distribution of reported visibility observations for calendar years 2020 – 2021.	46
5.3	Mean CRPS of post-processed, raw and climatological EUPPBench visibility forecasts for calendar year 2018 (a) and CRPSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.	47
5.4	Mean LogS of post-processed, raw and climatological EUPPBench visibility forecasts for calendar year 2018 (a) and LogSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.	48

5.5	CRPSS of POLR-L post-processed EUPPBench visibility forecasts for calendar year 2018 with respect to climatology (a) and LogSS of POLR-C and MLP-L approaches with respect to the POLR-L model (b) together with 95 % confidence intervals as functions of the lead time.	50
5.6	PIT histograms of post-processed, climatological and raw visibility EUPPBench forecasts for calendar year 2018 for lead times 6–30 h, 36–60 h, 66–90 h and 96–120 h.	51
5.7	Coverage (a) and average widths (b) of 90 % central prediction intervals of raw and post-processed EUPPBench visibility forecasts for calendar year 2018 as functions of the lead time. In panel (a) the ideal coverage is indicated by the horizontal dotted line.	52
5.8	RMSE of the mean EUPPBench visibility forecasts for calendar year 2018 (a) and difference in RMSE from climatology (b) as functions of the lead time.	53
5.9	Mean CRPS of climatological and post-processed visibility forecasts based on CAMS extended MLP and POLR models (a) and CRPSS of CAMS extended post-processed forecasts with respect to the corresponding MLP and POLR models based only on ECMWF predictions (b) as functions of the lead time. . . .	55
5.10	Mean LogS of climatological and post-processed visibility forecasts based on CAMS extended MLP and POLR models (a) and LogSS of CAMS extended post-processed forecasts with respect to the corresponding MLP and POLR models based only on ECMWF predictions (b) as functions of the lead time.	56
5.11	CRPSS (a) and LogSS (b) of CAMS extended MLP-C and POLR-L forecasts with respect to the corresponding models based only on ECMWF predictions together with 95 % confidence bounds as functions of the lead time.	56
5.12	Mean CRPS of raw, climatological and CAMS extended MLP-C and POLR-L forecasts censored from above at 10 km (a) and CRPSS of CAMS extended post-processed forecasts with respect to climatology (b) together with 95 % confidence bounds as functions of the lead time.	57
5.13	BSS of CAMS extended MLP-C and POLR-L forecasts with respect to climatology for the events where visibility is below 200 m (a), 500 m (b), 1000 m (c), and 3000 m (d) together with 95 % confidence bounds as functions of the lead time.	58

5.14	Mean ES of raw, climatological and post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models (a) and ESS of CAMS extended post-processed forecasts with respect to multivariate climatology (b) as functions of the lead time.	59
5.15	Mean VS of raw, climatological and post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models (a) and VSS of CAMS extended post-processed forecasts with respect to climatology (b) as functions of the lead time.	59
5.16	ESS (a) and VSS (b) post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models with respect to the corresponding forecasts based only on ECMWF predictions as functions of the lead time.	60
5.17	ESS (a,c) and VSS (b,d) of CAMS extended MLP-C ECC and POLR-L ECC multivariate forecasts with respect to multivariate climatology (a,b) and to the corresponding MLP-C ECC and POLR-L ECC forecasts based only on ECMWF predictions (c,d) together with 95 % confidence bounds as functions of the lead time.	61
5.18	Rank histograms of climatological and post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models.	62
5.19	Reliability indices of rank histograms of raw, climatological and post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models as functions of the lead time.	63
5.20	Boxplots of ECMWF visibility forecasts (logarithmic scale) for Fichtelberg mountain (Germany) for 0060 UTC (a), 1200 UTC (b), 1800 UTC (c), and 2400 UTC on the Christmas Eve of 2021 together with the corresponding CAMS forecasts and the median forecasts of the CAMS extended POLR-L model. The observed visibility is indicated by the horizontal dashed line. Values on the bottom line of panels (b) and (d) correspond to 0 m visibility.	65
6.1	Boxplots of the CRPSS of the EMOS models for T2M (a), V10 (b) and PPT24 (c) over the verification period with respect to the corresponding raw ensemble forecasts.	70

6.2	Boxplots of the mean ES over the verification period of the calibrated and raw T2M forecasts (a); of the ESS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance.	72
6.3	Boxplots of the mean VS over the verification period of the calibrated and raw T2M forecasts (a); of the VSS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance.	73
6.4	Boxplots of the DM test statistic investigating the significance of the difference in the mean ES (a) and mean VS (b) of Schaake shuffle methods based on historical observations from the rolling training period used in EMOS modelling with respect to the corresponding forecasts selecting dependence templates from the whole available past. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance.	74
6.5	Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw T2M forecasts.	75
6.6	Boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors of calibrated and raw T2M forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance.	75
6.7	Boxplots of the mean ES over the verification period of the calibrated and raw V10 forecasts (a); of the ESS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance.	76

6.8	Boxplots of the mean VS over the verification period of the calibrated and raw V10 forecasts (a); of the VSS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.	77
6.9	Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw V10 forecasts.	79
6.10	Boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors of calibrated and raw V10 forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.	79
6.11	Boxplots of the mean ES over the verification period of the calibrated and raw PPT24 forecasts (a); of the ESS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.	80
6.12	Boxplots of the mean VS over the verification period of the calibrated and raw PPT24 forecasts (a); of the VSS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.	81
6.13	Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw PPT24 forecasts.	82
6.14	Boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors of calibrated and raw PPT24 forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.	82

B.1	Mean CRPS of post-processed, raw and climatological visibility forecasts for calendar year 2021 (a) and CRPSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.	113
B.2	Mean LogS of post-processed, raw and climatological visibility forecasts for calendar year 2021 (a) and LogSS of post-processed forecasts with respect to the POLR-L model (b) as functions of the lead time.	114
B.3	CRPSS of POLR-L post-processed visibility forecasts for calendar year 2021 with respect to climatology (a) and LogSS of POLR-C and MLP-L approaches with respect to the POLR-L model (b) together with 95 % confidence intervals as functions of the lead time.	115
B.4	PIT histograms of post-processed, climatological and raw visibility forecasts for calendar year 2021 for lead times 6–60 h, 66–120 h, 126–180 h and 186–240 h.	116
B.5	Coverage (a) and average widths (b) of 90 % central prediction intervals of raw and post-processed visibility forecasts for calendar year 2021 as functions of the lead time. In panel (a), the ideal coverage is indicated by the horizontal dotted line.	117
B.6	RMSE of the mean forecasts for calendar year 2021 (a) and difference in RMSE from climatology (b) as functions of the lead time.	118
C.1	Boxplots of the DM test statistic investigating the significance of the difference in RDSS from the reference ECC-Q method for T2M (a), V10 (b) and PPT24 (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.	120

List of Tables

4.1	Variances of PIT values of post-processed forecasts of 10-m wind speed for lead times 1, 4, 7 and 10 days. The reference variance of the ideal standard uniform distribution is 0.083.	39
5.1	Overall mean CRPS/LogS of post-processed and climatological EUPPBench visibility forecasts for calendar year 2018 as proportion of the mean CRPS/LogS of the raw ECMWF ensemble.	48
5.2	Overall mean p -values of the α_{1234}^0 tests for uniformity of the PIT values for calendar year 2018.	52
B.1	Overall mean CRPS/LogS of post-processed and climatological visibility forecasts for calendar year 2021 as proportion of the mean CRPS/LogS of the raw ECMWF ensemble.	114
B.2	Overall mean p -values of the α_{1234}^0 tests for uniformity of the PIT values for calendar year 2021.	117

Introduction

Weather behaves chaotically. Minor deviations in the initial conditions of numerical weather prediction (NWP) models, which provide weather forecasts, might escalate quickly and significantly impact predictability. Moreover, from a single NWP output one cannot obtain information about the forecast uncertainty. However, the integration of ensemble prediction systems (EPSs) in operational use has revolutionized weather forecasting by enabling probabilistic predictions (Gneiting and Raftery, 2005). Each member of an ensemble forecast is produced from different runs of an NWP model, which can vary in their initial conditions and/or parameterizations (Bauer *et al.*, 2015; Buizza, 2018a). However, creating ensemble forecasts is far from simple, as the computational demand for generating such predictions increases both with the ensemble size and the forecast resolution. One of the leading weather services pioneering ensemble forecasting is the intergovernmental European Centre for Medium-Range Weather Forecasts (ECMWF), which in 30 years has significantly enhanced its operational Integrated Forecasting System (IFS). The horizontal resolution of the medium-range ensemble predictions has seen drastic improvements, initially at 120 km in 1992, it was reduced to 80 km in November 2000, to 32 km in January 2010, to 18 km by March 2016, and further refined to 9 km by June 2023¹.

While the advantages of probabilistic forecasts over their deterministic counterparts are clear, they also have drawbacks. Despite continuous improvements, the NWP models often contain systematic errors and may underestimate or overestimate actual observations, rendering them uncalibrated (Buizza, 2018b). In the early 2000s, the first statistical post-processing models emerged with the goal of mitigating these limitations of ensemble forecasts (Wilks, 2006). Among these are parametric approaches like ensemble model output statistics (EMOS; Gneiting *et al.*, 2005) and similar alternatives such as Bayesian model averaging (BMA; Raftery *et al.*, 2005) aim to map ensemble predictions onto a full predictive distribution, thereby addressing the inherent uncertainties and deficiencies associated with ensemble forecasts. Post-processing of ensemble

¹<https://www.ecmwf.int/en/about/who-we-are/history>

forecasts remains a hot topic to this day, with increasing popularity in the use of machine learning algorithms (Rasp and Lerch, 2018).

However, the post-processing methods mentioned above are usually individually applied to each prediction horizon, location, and time point. Consequently, dependencies between these elements may be overlooked during statistical calibration. Nevertheless, maintaining physical consistency across inter-variable, spatial, and temporal dependence structures is essential in numerous applications. The objective of multivariate post-processing is to incorporate dependence patterns into the process of ensemble calibration (Scheffzik and Möller, 2018). Methods for achieving this goal have existed since the early/mid 2010s, including the application of empirical (Scheffzik *et al.*, 2013) or parametric copulas (Möller *et al.*, 2013), or one could directly model the joint distributions as well (Baran and Möller, 2015); moreover, these methods can also be combined (Scheffzik, 2016).

One of the principal aims of this thesis is to conduct a thorough comparison and review of the performance of various multivariate post-processing procedures, as previous evaluations of such scope across multiple methods have primarily relied on simulated data (see e.g. Lerch *et al.*, 2020). This comparison is based on various multivariate verification metrics to provide a comprehensive assessment of the strengths and potential weaknesses of each method. Our focus is on the broadly applicable two-step methodologies, wherein ensemble predictions undergo initial separate post-processing for each variable, followed by the restoration of multivariate dependencies using copula functions in a subsequent step.

Another objective is the statistical post-processing of visibility ensemble forecasts, which are far less skillful compared to predictions of other weather variables. This generates a significant demand for developing efficient calibration techniques for this discrete quantity, as there are only a very few established ones. A related objective was to demonstrate that deterministic visibility forecasts from the Copernicus Atmosphere Monitoring Service (CAMS) could significantly enhance the success of the proposed post-processing methods. Additionally, we aimed to illustrate how these forecasts contribute to better preserving the spatial dependencies of visibility forecasts obtained through both univariate and multivariate calibration.

Finally, our last goal was to extend the clustering-based semi-local parameter estimation method proposed by Lerch and Baran (2017) with an interpolation step. The proposed approach was tested on global ECMWF 10-m wind speed forecasts to determine the parameters of the truncated normal EMOS model (Thorarinsdottir and Gneiting, 2010) at those locations where direct model training was impossible due to the lack of verifying observations.

The structure of the thesis is organized as follows to address the research objectives outlined earlier.

Chapter 1 provides a comprehensive overview of ensemble prediction systems, emphasizing their advantages and limitations, and transitioning to the rationale for post-processing. It introduces various variants of the parametric EMOS models used for calibrating different weather variables, along with state-of-the-art classification algorithms for post-processing discrete quantities.

In Chapter 2, we delve into the contemporary two-step multivariate post-processing methods we tested, explaining their precise operational mechanisms.

Chapter 3 covers widely used evaluation metrics in atmospheric sciences, which we applied extensively. This chapter also details the training process, spatial selection methods, and parameter estimation procedure.

Chapter 4 presents our motivation for extending the clustering-based semi-local parameter estimation method of Lerch and Baran (2017), the interpolation algorithms, a description of the wind speed data used in our case study, and the corresponding results (Baran and Lakatos, 2024).

Next, in Chapter 5, we discuss the statistical post-processing of visibility forecasts using data from ECMWF, integrating findings from both Baran and Lakatos (2023) and Lakatos and Baran (2024). This chapter begins with the calibration of forecasts using features derived solely from ECMWF data, followed by an examination of the impact of CAMS forecasts as additional covariates on the success of post-processing methods in both univariate and multivariate setups.

In Chapter 6, we focus on comparing multivariate post-processing methods, presenting the global ECMWF forecast-observation pairs used in our study, and showcasing our results (Lakatos *et al.*, 2023).

Finally, the dissertation concludes with Chapter 7, which offers a comprehensive summary of our findings and outlines potential directions for future research.

Chapter 1

Ensemble forecasting and univariate post-processing

1.1 Ensemble forecasts

NWP forecasts future atmospheric conditions by solving complex equations that describe atmospheric dynamics and physics, a concept first introduced by Lorenz (1963). Nowadays, as the computational demand for such tasks is much more manageable, the calculations are performed by a suite of programs running on supercomputers, collectively known as the NWP model. In the NWP model, the atmosphere is divided into a grid, with the intersection points referred to as grid points. Increasing the number of grid divisions results in a more detailed simulation. The model forecasts the future state of the atmosphere at each grid point based on the initial conditions. When multiple forecasts are generated by varying the initial conditions or using different models, the resulting set of predictions is known as an ensemble forecast. The general idea is that instead of relying on a single observed initial state, this method selects a finite set of initial states that are similar enough to the observed state to account for observational errors or inaccuracies. By applying a previously validated system of dynamic equations to each state in the ensemble, a range of future states is generated. This range can then be used to estimate the probability of specific events and calculate statistical measures like the ensemble mean and standard deviation for any given variable (Wilks, 2019, Chapter 8).

In ensemble forecasting, increasing the number of ensemble members enhances the system's ability to encompass a broader spectrum of potential outcomes with greater confidence. Likewise, higher spatial resolution, which dictates the granularity of grid cells or spatial units used in predictions, allows for more precise modelling of local conditions, thereby enhancing forecast ac-

curacy. One prominent weather center headquartered in the United Kingdom with sites in Bonn and Bologna is the European Centre for Medium-Range Weather Forecasts (ECMWF), which principally provides medium-range ensemble forecasts. Its Ensemble Prediction System (EPS) was initially deployed in June 1992, building upon theoretical work including that of Epstein (1969) and Leith (1974). These improvements include both a high-resolution deterministic forecast (HRES) and an ensemble forecast (ENS). The ensemble forecast consists of 51 members: one control forecast (CNTL) and 50 forecasts with slightly different initial conditions and perturbations to model physics. All forecasts have a horizontal resolution of approximately 9 km. Combined, they provide detailed insights into the weather evolution up to 15 days in advance (ECMWF, 2024b). Beside the ECMWF IFS, numerous weather forecasting centres have played a crucial role in improving the accuracy and reliability of weather predictions by accounting for uncertainties and providing probabilistic forecasts. NOAA’s Global Ensemble Forecast System (GEFS; Zhou *et al.*, 2022), initiated in December 1992, and the UK Met Office’s Global and Regional Ensemble Prediction System (MOGREPS; Bowler *et al.*, 2008), launched in the early 2000s, are notable examples, while Germany’s ICON model has introduced the ICON Ensemble Prediction System to further refine probabilistic forecasting capabilities. Additional significant forecasting systems outside Europe and America include Japan’s JMA-EPS, China’s GRAPES, Australia’s ACCESS, South Korea’s KMA systems, and India’s IMD models. Finally, the AROME-EPS system of the HungaroMet, currently produces 11 members. This model is primarily used for ultra-short-term forecasts, with a horizontal spatial resolution of 2.5 km (Jávorné Radnóczy *et al.*, 2020).

In addition to traditional NWP models, there is an increasing inclination towards developing data-driven machine learning-based approaches in weather prediction. Huawei’s Pangu-Weather stands out as an AI-driven model that has exhibited superior precision compared to traditional numerical weather forecast methods. It delivers stronger deterministic forecast results across various variables using reanalysis data, which integrates past short-range weather forecasts and observations. This model shows promising performance when directly compared with the ECMWF IFS (Bi *et al.*, 2023). Another noteworthy AI-based model is NVIDIA’s FourCastNet (Kurth *et al.*, 2023), which achieves comparable accuracy to the ECMWF IFS while operating significantly faster and with higher energy efficiency than traditional methods. Additionally, DeepMind’s GraphCast (Lam *et al.*, 2023) have demonstrated high performance, offering promising results in terms of speed and accuracy. By the end of 2023, ECMWF launched the alpha version of their data-driven Artificial Intelligence/Integrated Forecasting System (AIFS) model (Lang *et al.*, 2024). Since then, they have introduced the first version of AIFS to generate ensemble forecasts. Furthermore, with forecasts reaching up to 15 days, Google Research’s

GenCast (Price *et al.*, 2025) outperforms the ECMWF ENS by providing more accurate predictions for both typical weather patterns and extreme events.

However, despite continuous improvements, the drawback of ensemble forecasts remains rooted mainly in the NWP model’s sensitivity to initial conditions, which can lead to biased, under-dispersive, or over-dispersive predictions, which calls for some form of post-processing. Among the various approaches for post-processing ensemble forecasts, these can be categorized into parametric, non-parametric, and machine learning-based models. Non-parametric techniques, such as quantile regression forests (utilized by Taillardat *et al.* (2016)), have gained considerable attention due to their ability to capture complex relationships without assuming a specific parametric structure for the underlying distribution, allowing for the estimation of desired quantiles. Additionally, gradient boosting machines (Friedman, 2001) consist of ensembles of decision trees that are grown sequentially, using information from previously grown trees. In this process, each decision tree is trained on a modified version of the original training set, focusing on regions where previous model iterations provided poor predictions. This model was used, for example, by Baran *et al.* (2021) for calibrating total cloud cover forecasts, but since tree-based methods like random forests and gradient boosting machines were outperformed by several other approaches, we decided against its utilization in discrete post-processing.

1.2 Ensemble model output statistics

Ensemble model output statistics (EMOS; Gneiting *et al.*, 2005) is a highly useful yet simple regression-based parametric post-processing method which nowadays commonly used as a reference approach to ensemble calibration (see e.g. Vannitsem *et al.*, 2021; Schultz and Lerch, 2022). EMOS maps ensemble forecasts of a weather quantity onto a predictive distribution represented by a single parametric law, the type of which and/or the link functions connecting its parameters to the forecasts depend on the nature of the variable being calibrated.

The location/mean of these predictive distributions is usually determined as an affine function of the ensemble members, reflecting the relative influence of each member within the ensemble forecast, whereas the scale/standard deviation could be an affine function of the ensemble mean, ensemble variance or the ensemble mean difference. Following the optimum score principle of Gneiting and Raftery (2007), the estimation of the hyper-parameters of such a predictive distribution is mostly carried out by minimizing the mean of a proper verification score (see Section 3.1) over the training data comprising past forecast-observation pairs. EMOS predictive distributions are now applicable across a wide range of meteorological variables, such as temperature (Gneiting *et al.*, 2005; Taillardat, 2021), wind speed (Thorarinsdottir and Gneiting, 2010; Lerch

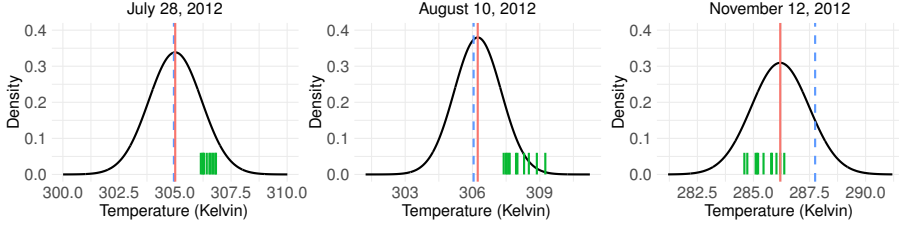


Figure 1.1: Calibrated normal EMOS probability density functions for the HungaroMet’s 42 h temperature forecasts, for Debrecen Airport for 3 different dates, illustrating the EMOS median (red solid line), raw ensemble (green segments) and observation (blue dashed line).

and Thorarinsdottir, 2013; Baran and Lerch, 2015; Szabó *et al.*, 2023), precipitation accumulation (Scheuerer, 2014; Baran and Nemoda, 2016), various heat indices (Baran *et al.*, 2020) or solar irradiance (Schultz *et al.*, 2021).

In most cases, the calibrated predictive distribution more accurately captures potential observations compared to the ensemble forecasts alone. The EMOS median typically aligns much closer to the actual observations, highlighting the effectiveness of bias correction. Figure 1.1 illustrates such cases with temperature forecasts, produced by the operational EPS of the HungaroMet.

In the next subsections, we introduce EMOS models associated with the weather quantities we have analyzed. In what follows, let us denote by $f_1, f_2, \dots, f_K$ a K -member ensemble forecast, for a given location, time point and prediction horizon, where $\bar{f}$ represents the ensemble mean, and S^2 and MD the ensemble variance and ensemble mean difference, respectively, defined as

$$S^2 := \frac{1}{K-1} \sum_{k=1}^K (f_k - \bar{f})^2 \quad \text{and} \quad \text{MD} := \frac{1}{K^2} \sum_{k,\ell=1}^K |f_k - f_\ell|.$$

If an ensemble consists of M ensemble members organized into K exchangeable groups, where each k th group contains $M_k \geq 1$ ensemble members (with the total summing up to M), we denote its mean as $\bar{f}_k$. The term “exchangeable” refers to the statistical indistinguishability of the members within each group.

In our case studies we will work with the ECMWF medium-range ensemble forecasts, that comprise a total of 52 members. Among these, the HRES member provides a detailed forecast using a finer grid. The CNTL member utilizes the actual atmospheric state as initial conditions for its generation as the NWP input. Additionally, the system includes 50 perturbed forecasts derived from variations in initial conditions. These latter 50 forecasts can be considered statistically interchangeable. Thus, the HRES, CNTL, and the 50-member ENS

forecasts form three distinct groups within the ECMWF ensemble.

In the following, let f_{CNTL} , f_{HRES} and $\bar{f}_{ENS}$ represent the ECMWF control- and high-resolution forecasts, and the mean of the 50 exchangeable forecasts, respectively.

1.2.1 Normal EMOS model

The normal distribution and its generalizations (such as skewed normal, mixture of normals, etc., see, e.g., Gneiting *et al.* (2005); Raftery *et al.* (2005); Rasp and Lerch (2018)) are widely used for modelling temperature, and another example for a quantity modeled by a normal distribution is pressure. Generally, the EMOS predictive distribution proposed by Gneiting *et al.* (2005) specifies the mean μ and variance σ^2 as

$$\mu = a_0 + a_1^2 f_1 + a_2^2 f_2 + \cdots + a_{K-1}^2 f_{K-1} + a_K^2 f_K, \quad \sigma^2 = b_0^2 + b_1^2 S^2,$$

where in the case of having K exchangeable groups with group means $\bar{f}_i$ one has

$$\mu = a_0 + \sum_{i=1}^K a_i^2 \bar{f}_i.$$

where $a_0, a_1, a_2, \dots, a_K, b_0, b_1 \in \mathbb{R}$ are the model parameters (Gneiting, 2014).

In Chapter 6 of this thesis, our results on the calibration of ECMWF 2-meter temperature forecasts will be discussed. In that study, we utilized a variant of the normal EMOS model designed to capture the seasonal variations present in the temperature data. To account for seasonal variations in temperature, a more complex model for the mean of the predictive distribution is suggested by Scheuerer and Büermann (2014). Consider the regression model

$$y_t = c_0 + c_1 \sin\left(\frac{2\pi t}{365}\right) + c_2 \cos\left(\frac{2\pi t}{365}\right) + \varepsilon_t, \quad t = 1, 2, \dots, n, \quad (1.2.1)$$

where y_t represents either temperature observations for a given location or functionals of the corresponding ECMWF ensemble forecast with a given lead time ℓ , namely the high-resolution member f_{HRES} , the control run f_{CNTL} , and the mean of the exchangeable forecasts $\bar{f}_{ENS}$ from a training period of length n . The phase parameters vary across different training periods, resulting in the evolution of c_1 and c_2 throughout the calendar year. With the help of model (1.2.1), one can calculate the ℓ -step ahead predictions $\hat{y}$ and $\hat{f}_{HRES}$, $\hat{f}_{CNTL}$, $\hat{f}_{ENS}$ of the observations and the corresponding functionals of the ensemble, respectively, and obtain a Gaussian predictive distribution with parameters (1.2.2) and (1.2.3):

$$\mu = \hat{y} + a_1^2 (f_{HRES} - \hat{f}_{HRES}) + a_2^2 (f_{CNTL} - \hat{f}_{CNTL}) + a_3^2 (\bar{f}_{ENS} - \hat{f}_{ENS}) \quad (1.2.2)$$

$$\sigma^2 = b_0^2 + b_1^2 S^2, \quad (1.2.3)$$

where

$$S^2 := \frac{1}{52} \sum_{k=1}^{52} (f_k - \bar{f})^2 \quad \text{with} \quad \bar{f} := \frac{1}{52} \sum_{k=1}^{52} f_k.$$

1.2.2 Truncated normal EMOS model

To model wind speed, non-negative and skewed distributions are necessary. In this context, we consider a normal distribution $\mathcal{N}_0^\infty(\mu, \sigma^2)$ with location parameter μ and scale parameter $\sigma > 0$, truncated from the left at zero, which is applied in the EMOS model of Thorarinsdottir and Gneiting (2010). For the ECMWF ensemble, the location and scale parameters of the predictive distribution simplify to:

$$\mu = a_0 + a_1^2 f_{\text{HRES}} + a_2^2 f_{\text{CNTL}} + a_3^2 \bar{f}_{\text{ENS}} \quad \text{and} \quad \sigma^2 = b_0^2 + b_1^2 S^2. \quad (1.2.4)$$

Instead of S^2 , another common choice is the use of the ensemble mean difference as a dispersion measure, which is:

$$\text{MD}_f := \frac{1}{52^2} \sum_{k,\ell=1}^{52} |f_k - f_\ell|. \quad (1.2.5)$$

As a natural transformation (see e.g., Haslett and Raftery, 1989), one could also model the square root of wind speed, as demonstrated by Hemri *et al.* (2014) and Lakatos *et al.* (2023). The model parameters $a_0, a_1, a_2, a_3, b_0, b_1 \in \mathbb{R}$ can again be estimated using the optimum score estimation principle.

1.2.3 Censored generalized extreme value EMOS model

Calibrating precipitation forecasts statistically is more difficult compared to post-processing temperature or wind speed forecasts due to its discrete-continuous nature, along with the great likelihood of zero precipitation occurring. In line with Hemri *et al.* (2014), in Chapter 6 we model 24-hour precipitation accumulation using the EMOS model proposed by Scheuerer (2014), which employs a generalized extreme value distribution that is left-censored at zero. Location μ and scale σ are linked to the raw forecasts via

$$\mu = a_0 + a_1^2 f_{\text{HRES}} + a_2^2 f_{\text{CNTL}} + a_3^2 \bar{f}_{\text{ENS}} + a_4^2 \pi_0 \quad \text{and} \quad \sigma^2 = b_0^2 + b_1^2 \text{MD}_f, \quad (1.2.6)$$

where $a_0, a_1, a_2, a_3, a_4, b_0, b_1 \in \mathbb{R}$ and

$$\pi_0 := \frac{1}{52} \sum_{k=1}^{52} \mathbb{I}_{\{f_k=0\}}$$

is the proportion of ensemble members predicting zero precipitation, whereas shape ξ is kept fixed ($\xi = 0.2$).

1.3 Discrete post-processing

When dealing with discrete weather variables, the application of parametric EMOS models becomes impractical. An illustrative example is the calibration of total cloud cover as a discrete variable, where ensemble forecasts estimate the fraction of sky covered by clouds, expressed in octas. This means that observations of total cloud cover can only take on nine distinct values. Another significant example is visibility, which is also reported in discrete values; however, in this case, the number of distinct observable values is much larger. In such instances, probabilistic weather forecasting shifts to a classification task, where total cloud cover has 9 observation categories and visibility has 84 distinct observable values (based on the reporting scheme of the World Meteorological Organization). These values need to be predicted in terms of their probabilities rather than being fitted into a complete parametric distribution, see, e.g., Baran and Lakatos (2023).

1.3.1 Proportional odds logistic regression

As mentioned, in the context of discrete weather quantities, post-processing simplifies to a classification problem where multiclass logistic regression (MLR) serves as a prevalent and effective tool. For a classification problem with C distinct categories and an M -dimensional feature vector $\mathbf{x}$, the MLR model requires the estimation of $(C - 1)(M + 1)$ parameters from training data to ensure numerical stability during estimation. This necessitates a substantial training dataset comprising historical forecast-observation pairs.

An alternative, more streamlined approach is the proportional odds logistic regression model (McCullagh, 1980), tailored for ordered data such as visibility observations. Here, the conditional cumulative distribution of the discrete weather quantity Y with range $\mathcal{Y} = \{y_1, y_2, \dots, y_C\}$ given an M -dimensional feature vector $\mathbf{x}$ is formulated as:

$$P(Y \leq y_k | \mathbf{x}) = \frac{e^{\mathcal{L}_k(\mathbf{x})}}{1 + e^{\mathcal{L}_k(\mathbf{x})}}, \quad \text{where} \quad \mathcal{L}_k(\mathbf{x}) = \alpha_k - \mathbf{x}^\top \boldsymbol{\beta}, \quad k = 1, 2, \dots, C,$$

with $\alpha_k \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^M$, and assuming $\alpha_1 < \alpha_2 < \dots < \alpha_C$. Therefore, the POLR model involves $C + M$ unknown parameters.

1.3.2 Multilayer perceptron

For both continuous and discrete weather quantities, the application of neural networks has increased in popularity for calibration. For the latter, instead of a parametric approach, one can use a classical feedforward multilayer perceptron (MLP) neural network for classification. A classical MLP consists of multiple layers and neurons or nodes, each of which is a transformed (by an

activation function) weighted sum of the node values from the previous layer plus an additional bias term. The features are "fed" to the input layer, and the predictions of the distribution of the various classes are made on the output layer. The number of hidden layers and neurons in them are tuning parameters of the network and provide the level of abstraction. For additional information, we refer to Goodfellow *et al.* (2016).

Chapter 2

Multivariate post-processing

Independent univariate calibration of ensemble forecasts for each ℓ forecast horizon and station s does not account for the temporal dependencies between predictions initialized at the same time point or the spatial correlations between observation stations. These dependencies are restored in a second step with the help of the approaches described below, which are based on empirical copulas. In what follows, let d be a general notion representing a dimension ℓ or s of an ensemble forecast and let $\mathbf{f}^{(d)} = (f_1^{(d)}, f_2^{(d)}, \dots, f_M^{(d)})$ denote an M -member ensemble forecast for dimension d .

In general, a copula is a multivariate cumulative distribution function (CDF) with standard uniform marginals. According to Sklar’s theorem (Sklar, 1959), any D -dimensional CDF H with marginal CDF-s $F^{(d)}$, $d = 1, 2, \dots, D$, can be decomposed as

$$H(x_1, x_2, \dots, x_D) = C(F^{(1)}(x_1), F^{(2)}(x_2), \dots, F^{(D)}(x_D)), \quad x_1, x_2, \dots, x_D \in \mathbb{R},$$

where C is a D -dimensional copula representing the dependencies between the marginals.

Here we focus on non-parametric methods where marginals are represented by empirical CDFs $\hat{F}^{(d)}$ of samples drawn independently from the corresponding predictive distributions, and C is an empirical copula (Rüschendorf, 2009) providing a “dependence template” derived from a given discrete dataset. The learned dependence structure is then imposed on the post-processed forecasts by rearranging the marginal samples according to the specified template (see e.g. Wilks, 2019, Section 8.4.2).

2.1 Ensemble Copula Coupling

In the ensemble copula coupling (ECC) approach the dependence template is obtained from the corresponding raw ensemble forecasts. The method, introduced by Schefzik *et al.* (2013), consists of the following two simple steps.

1. For each dimension $d = 1, 2, \dots, D$, generate a sample $\hat{f}_1^{(d)}, \hat{f}_2^{(d)}, \dots, \hat{f}_M^{(d)}$ of the same size M as the raw ensemble from the calibrated marginal predictive distribution $F^{(d)}$, which is assumed to be arranged in ascending order.
2. Consider permutations $\pi_d = (\pi_d(1), \pi_d(2), \dots, \pi_d(M))$ of $\{1, 2, \dots, M\}$ induced by the rank order structure of the raw ensemble $f_1^{(d)}, f_2^{(d)}, \dots, f_M^{(d)}$, that is $\pi_d(m) := \text{rank}(f_m^{(d)})$ with ties resolved at random. The ECC calibrated sample $\tilde{f}_1^{(d)}, \tilde{f}_2^{(d)}, \dots, \tilde{f}_M^{(d)}$ for dimension d is obtained by rearranging the sample generated in step 1 according to permutation π_d , that is

$$\tilde{f}_m^{(d)} := \hat{f}_{\pi_d(m)}^{(d)}, \quad m = 1, 2, \dots, M, \quad d = 1, 2, \dots, D.$$

Similar to Lerch *et al.* (2020), we consider three different ECC variants depending on the sampling method in step 1. **ECC-R** refers to random sampling from the predictive distribution $F^{(d)}$ and arranging the sample in ascending order, whereas in **ECC-Q** one considers equidistant quantiles of $F^{(d)}$, that is

$$\tilde{f}_m^{(d)} := (F^{(d)})^{-1} \left(\frac{m}{M+1} \right), \quad m = 1, 2, \dots, M, \quad d = 1, 2, \dots, D.$$

Finally, we also apply the stratified sampling approach of Hu *et al.* (2016) and refer to the corresponding ECC variant as **ECC-S**. In this approach

$$\tilde{f}_m^{(d)} := (F^{(d)})^{-1}(u_m), \quad m = 1, 2, \dots, M, \quad d = 1, 2, \dots, D,$$

where u_m is a random draw from a uniform distribution on $]\frac{m-1}{M}, \frac{m}{M}]$, $m = 1, 2, \dots, M$.

2.2 Dual Ensemble Copula Coupling

The dual ensemble copula coupling (dECC) method of Ben Bouallègue *et al.* (2016) combines the structure of the raw ensemble forecast with the estimated forecast error autocorrelation, and proceeds as follows.

1. Apply ECC-Q to generate an initial post-processed multivariate ensemble forecast $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \dots, \tilde{\mathbf{f}}_M$ with $\tilde{\mathbf{f}}_m = (\tilde{f}_m^{(1)}, \tilde{f}_m^{(2)}, \dots, \tilde{f}_m^{(D)})^\top$, $m = 1, 2, \dots, M$.

2. With the help of the estimated $D \times D$ autocorrelation matrix $\widehat{\Sigma}_e$ of the forecast error of the ensemble mean generate a correction term

$$\mathbf{c}_m := \widehat{\Sigma}_e^{1/2} (\tilde{\mathbf{f}}_m - \mathbf{f}_m), \quad m = 1, 2, \dots, M,$$

where $\mathbf{f}_m = (f_m^{(1)}, f_m^{(2)}, \dots, f_m^{(D)})^\top$ denotes the m th raw multivariate forecast. The estimates of the error correlations can be obtained e.g. from the training data for the univariate post-processing at the different forecast horizons.

3. Derive the adjusted multivariate ensemble $\check{\mathbf{f}}_1, \check{\mathbf{f}}_2, \dots, \check{\mathbf{f}}_M$, where $\check{\mathbf{f}}_m := \mathbf{f}_m + \mathbf{c}_m$, $m = 1, 2, \dots, M$.
4. Apply ECC-Q again; however, this time using the rank order structure of the adjusted ensemble of step 3 for rearranging the samples generated from the calibrated univariate predictive distributions.

2.3 Schaafe Shuffle

By contrast to ECC, the Schaafe shuffle (SSh; Clark *et al.*, 2004) determines a dependence template based on past observations rather than the raw ensemble predictions. Samples drawn from the calibrated univariate predictive distributions are thus rearranged in the rank order structure of randomly selected historical observation vectors of length D . Again, as historical data one can consider the training data for the univariate post-processing. The SSh approach allows for generating post-processed forecasts of any size, provided one has a long enough historical climatological record. However, to ensure a fair comparison with the (d)ECC methods, we restrict the sample size to the size of the raw ensemble. Similar to the ECC described in Section 2.1, three different sampling methods from the predictive distributions are considered; the corresponding SSh variants are referred to as **SSh-R** (random sample), **SSh-Q** (equidistant quantiles) and **SSh-S** (stratified sample). In addition to the standard SSh, we further apply two more recently proposed variants that utilize more advanced procedures to select past observations for determining the dependence template.

2.4 Minimum Divergence Schaafe Shuffle

The minimum divergence Schaafe shuffle (mdSSh; Scheuerer *et al.*, 2017) provides a more sophisticated method for selecting historical observation vectors used as a dependence template.

1. For each dimension d , calculate the 99% central prediction interval of the corresponding marginal predictive distribution $F^{(d)}$, $d = 1, 2, \dots, D$.
2. From the historical climatological record keep those D -dimensional observation vectors (e.g. observation trajectories, if the dimension d refers to the forecast horizon of predictions initialized at the same time as in Lakatos *et al.* (2023)), where the corresponding central prediction intervals of step 1 contain at least k observations. Threshold k is chosen to retain at least $K \geq M$ forecast vectors.
3. Select randomly M observation vectors from the K remaining after step 2.

For the discrete-continuous marginal predictive distributions of 24-hour precipitation accumulation forecasts used in our work (see Chapter 6), Scheuerer *et al.* (2017) suggests to replace step 3 by a more complex method to reduce the number of selected forecast vectors from K to the required M . This approach is based on selection of a M -subset of the set of forecast vectors of step 2, which minimizes the total divergence of the calibrated predictive CDFs and the empirical CDFs of the corresponding observations (see e.g. Thorarinsdottir *et al.*, 2013) for all dimensions and forecast cases. For more details and a computationally feasible algorithm we refer to Scheuerer *et al.* (2017).

2.5 Similarity-based Schaake Shuffle

In contrast to the mdSSh, where observation vectors used as dependence templates are selected on the basis of their consistency with the corresponding calibrated predictive distributions, the similarity-based Schaake shuffle (simSSh; Schefzik, 2016) looks for historical forecast vectors where the corresponding ensemble forecasts are the most similar to the actual ones.

1. For a specific date calculate the mean $\bar{f}^{(d)}$ and variance $S^{2(d)}$ of the corresponding ensemble forecast $\mathbf{f}^{(d)}$ for each dimension $d = 1, 2, \dots, D$.
2. For each date τ in the historical data set compute similarity

$$\Delta^\tau := \sqrt{\sum_{d=1}^D \left(\bar{f}^{(d)} - \bar{f}_\tau^{(d)} \right)^2 + \frac{1}{D} \sum_{d=1}^D \left(S^{2(d)} - S_\tau^{2(d)} \right)^2},$$

where $\bar{f}_\tau^{(d)}$ and $S_\tau^{2(d)}$ denote the mean and variance of the ensemble forecast $\mathbf{f}_\tau^{(d)}$ at dimension d corresponding to date τ , respectively.

3. Choose dates $\tau_1, \tau_2, \dots, \tau_M$ resulting in the highest similarity to the actual forecasts, that is where $\Delta^{\tau_1}, \Delta^{\tau_2}, \dots, \Delta^{\tau_M}$ are the smallest among all similarities computed in step 2. The dependence template is given by historic observations $y_{\tau_m}^{(d)}$ corresponding to ensemble forecasts $\mathbf{f}_{\tau_m}^{(d)}$, $m = 1, 2, \dots, M$, $d = 1, 2, \dots, D$.

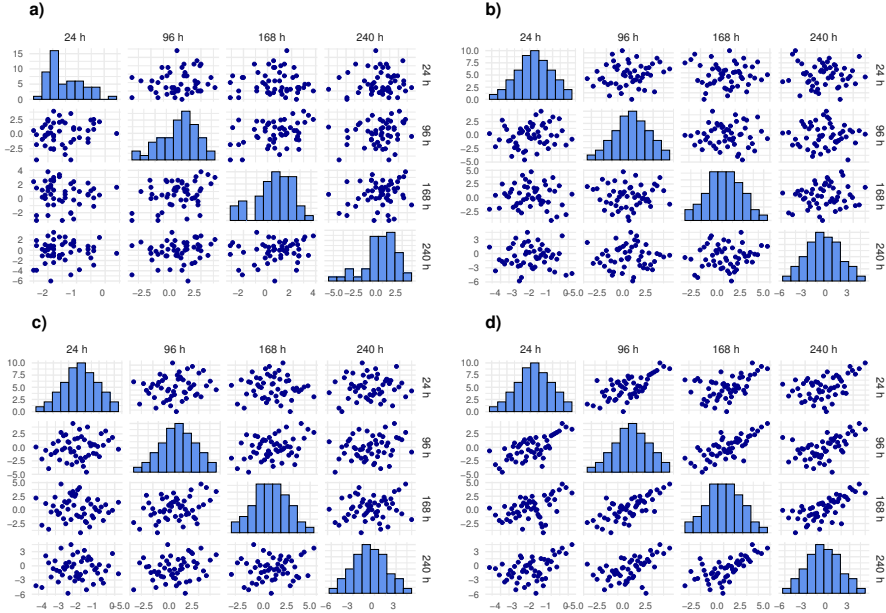


Figure 2.1: The effect of multivariate post-processing: 24, 96, 168, and 240 h ahead ECMWF ensemble forecasts for temperature initialized at 18 November 2006; a) raw ensemble forecasts b) after individual EMOS postprocessing (losing dependence structures) c) after ECC, and d) Schaake shuffle postprocessing.

2.6 Assumption of independence

In our case studies, we also investigate the forecast skill of naive multivariate forecasts obtained from the univariate post-processing methods, without accounting for temporal, spatial, or intervariable dependencies. This means that forecasts $\hat{f}_k^{(d)}$, $m = 1, 2, \dots, M$, $d = 1, 2, \dots, D$, drawn from the corresponding calibrated predictive distributions $F^{(d)}$ are simply combined into D -dimensional ensemble forecasts $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_M$ with $\hat{\mathbf{f}}_m = (\hat{f}_m^{(1)}, \hat{f}_m^{(2)}, \dots, \hat{f}_m^{(D)})^\top$,

$m = 1, 2, \dots, M$. These ‘independent’ forecasts can be derived using random sampling from the univariate predictive distributions; however, considering equidistant quantiles or stratified sampling is also a popular choice.

Figure 2.1 illustrates the effect of univariate and multivariate post-processing on 24-, 96-, 168-, and 240-hour ECMWF temperature forecasts. The scatter plots show the pairwise relationships between forecasts, while the marginal distributions are visualized along the main diagonals. As seen, the univariate post-processing successfully calibrates the normally distributed temperature forecasts, but the correlations between different forecast horizons are mostly lost. The application of ECC largely restores these correlations, whereas the Schaake shuffle method slightly overestimates the correlations in this case.

2.7 Operational implementation of multivariate post-processing

ECC and the SSh have been explored in research contexts in Austria, and within ECMWF’s ensemble predictions, but their operational implementation remains limited in many regions (Perrone *et al.*, 2020). However, ECC is operationally used in France, where it has been applied to enhance ensemble forecasting systems, particularly in improving precipitation and temperature predictions (Taillardat and Mestre, 2020). The specific reasons for its limited use in other regions are not publicly disclosed, and research into the potential benefits of these methods is ongoing both within meteorological organizations and the scientific community. Naturally, major changes to existing forecasting workflows are only justified if they lead to significant improvements in forecast skill. For example, ECMWF recently investigated the potential enhancement of univariate grid-point post-processed rainfall forecasts using multivariate methods. Overall, it has been observed that multivariate methods perform better than univariate approaches, and additional improvements could reduce forecast production costs, strike a balance between operational feasibility and accuracy, to make the multivariate approach more appealing for long-term use (Pillosu *et al.*, 2024).

Chapter 3

Forecast evaluation metrics and training data selection

Given that forecasts inherently involve uncertainty, accurately assessing it is crucial for informed decision-making. Scoring rules are used to evaluate probabilistic forecasts by assigning a numerical score that considers both the predicted probability distribution and the actual event or value that occurs. In general, a scoring rule is deemed proper if the forecaster achieves the highest/lowest expected score (depending on the orientation of the score) by providing the probabilistic forecast F , which matches the true distribution, rather than any alternative forecast $G \neq F$. It is strictly proper if this maximum score is uniquely attained with the forecast F (Gneiting and Raftery, 2007). Examples of scoring rules for probabilistic forecasts expressed as predictive densities include the logarithmic and the continuous ranked probability score (CRPS), which are discussed in more detail in Section 3.1. The strictly proper CRPS is now established as a standard metric for evaluating probabilistic forecasts, serves as the objective function for EMOS calibration or is increasingly used as the cost function in machine learning algorithms, due to its appealing characteristics.

The subsequent subsections will detail the scoring rules employed and outline the process for selecting training data.

3.1 Verification metrics

In the evaluation of the predictive performance of probabilistic forecasts, Gneiting *et al.* (2007) suggest the principle of “maximizing the sharpness of the predictive distributions subject to calibration”. Sharpness is the property of the predictive distribution only and measures its concentration. For instance, for a

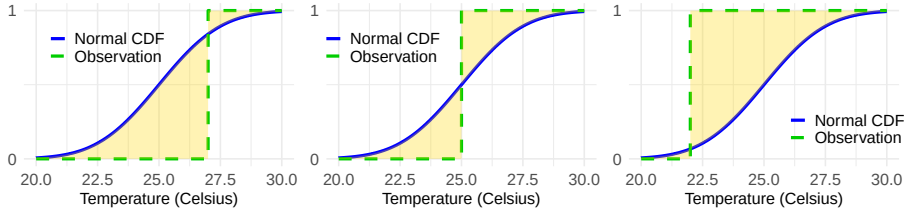


Figure 3.1: Illustration of the continuous ranked probability score (CRPS, gold area) through three simulated forecast cases (normal CDFs) and their corresponding observations

given $\alpha \in]0, 1[$ specifying the confidence level, one can investigate the average width of the $(1 - \alpha) \times 100\%$ central prediction intervals, that is intervals between the lower and upper $\alpha/2$ quantiles of the predictive distribution. The smaller this average width (reported in the same units as the weather quantity at hand), the more concentrated the predictive distribution. In contrast, for a calibrated probabilistic forecast predictive distributions and corresponding observations should be statistically consistent (predicted probability of an event should match the relative frequency of observing it), so calibration is their joint property. Considering again the above central prediction intervals, for a calibrated forecast their coverage (proportion of observations located in the corresponding central prediction intervals) should be around $(1 - \alpha) \times 100\%$ (see e.g. Gneiting and Raftery, 2007). Note that by choosing α to match the nominal coverage of $(M - 1)/(M + 1)100\%$ of the ensemble range of a M -member ensemble forecast, one can provide a fair comparison with the raw predictions.

The logarithmic score (LogS; Good, 1952), which allows for infinite scores, for a probabilistic forecast provided in a form of a probability density function (PDF) f can be expressed as

$$\text{LogS}(f, y) := -\log(f(y)).$$

Besides LogS, one of the most popular proper scoring rules in the atmospheric sciences assessing simultaneously both calibration and sharpness is the continuous ranked probability score (CRPS; Wilks, 2019, Section 9.5.1). For a predictive CDF $F(x)$ and an observation $y \in \mathbb{R}$, the CRPS is defined as

$$\text{CRPS}(F, y) := \int_{-\infty}^{\infty} (F(x) - \mathbb{I}_{\{x \geq y\}})^2 dx = \mathbb{E}|X - y| - \frac{1}{2}\mathbb{E}|X - X'|, \quad (3.1.1)$$

where $\mathbb{I}_H$ denotes the indicator of a set H , whereas X and X' are independent random variables with CDF F and finite first moment. The CRPS

is a negatively oriented score, that is the smaller the better, and this scoring rule serves as loss function in parameter estimation of normal, truncated normal and censored generalized extreme value EMOS models for calibration of temperature, 10-meter wind-speed and 24-hour precipitation accumulation ensemble forecasts, respectively. For these distributions the CRPS can be obtained in closed form (for the corresponding formulae see e.g. Jordan *et al.*, 2019), allowing a computationally efficient estimation process. The illustration in Figure 3.1 demonstrates the accuracy of normally distributed probabilistic forecasts based on the CRPS by comparing the predicted cumulative distribution function with the observed values, through three simulated forecast cases.

Additionally, to evaluate predictions for high-impact events, one can also consider the threshold-weighted continuous ranked probability score (twCRPS; Gneiting and Ranjan, 2011)

$$\text{twCRPS}(F, y) := \int_{-\infty}^{\infty} [F(x) - \mathbb{I}_{\{x \geq y\}}]^2 \omega(x) dx, \quad (3.1.2)$$

where $\omega(y) \geq 0$ is a weight function. Observations of a weather quantity with predictive distribution F exceeding a given threshold r can be addressed by setting $\omega(y) = \mathbb{I}_{\{y \geq r\}}$, whereas the choice $\omega(y) \equiv 1$ results in the traditional CRPS (3.1.1).

The Brier score (BS; Wilks, 2019, Section 9.4.2) can be employed for binary outcomes, such as when the observed weather quantity y doesn't exceed a given threshold x . Given a predictive CDF $F(x)$, the Brier score is computed as:

$$\text{BS}(F, y; x) := (F(x) - \mathbb{I}_{\{x \geq y\}})^2$$

The CRPS is then derived by integrating the BS over all possible threshold values.

Calibration of univariate ensemble forecasts can also be diagnosed with the help of verification rank histograms displaying the ranks of observations with respect to the ensemble forecasts (see e.g. Wilks, 2019, Section 9.7.1). For a properly calibrated M -member ensemble the ranks follow a uniform distribution on $\{1, 2, \dots, M+1\}$, and the deviation from uniformity can be quantified by the reliability index

$$\text{RI} := \sum_{r=1}^{M+1} \left| \rho_r - \frac{1}{M+1} \right|, \quad (3.1.3)$$

where ρ_r is the relative frequency of rank r (Delle Monache *et al.*, 2006).

In the case of probabilistic forecasts given in the form of predictive distributions, a similar role is played by the probability integral transform (PIT; Wilks, 2019, Section 9.5.4) histogram. PIT is the predictive CDF evaluated

at the verifying observation, which for a well calibrated probabilistic forecast follows a standard uniform law. Similar to the verification rank histograms, biased forecasts result in triangular shapes, whereas U-shaped and hump-shaped PIT histograms refer to under- and overdispersion, respectively.

In the simplified scenario where a single prediction represents the possible future outcome, the most common metrics for assessing fit are the mean absolute error (MAE), as the median minimizes this score (Gneiting, 2011) and the root mean squared error (RMSE) optimized by the mean.

A multivariate extension of the CRPS is the energy score (ES; Gneiting and Raftery, 2007). Given a D -dimensional CDF F and vector $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(D)})^\top$, the ES is defined as

$$\text{ES}(F, \mathbf{y}) := \mathbb{E} \|\mathbf{X} - \mathbf{y}\| - \frac{1}{2} \mathbb{E} \|\mathbf{X} - \mathbf{X}'\|, \quad (3.1.4)$$

where $\|\cdot\|$ denotes the Euclidean distance and, similar to the univariate case (3.1.1), $\mathbf{X}$ and $\mathbf{X}'$ are independent random vectors having distribution F . For a forecast ensemble $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_M$ one should consider the empirical CDF F_M (Gneiting *et al.*, 2008), which reduces (3.1.4) to the ensemble energy score

$$\text{ES}(F_M, \mathbf{y}) = \frac{1}{M} \sum_{j=1}^M \|\mathbf{f}_j - \mathbf{y}\| - \frac{1}{2M^2} \sum_{j=1}^M \sum_{m=1}^M \|\mathbf{f}_j - \mathbf{f}_m\|. \quad (3.1.5)$$

A more recently introduced multivariate proper scoring rule is the (ensemble) variogram score of order p (VS^p ; Scheuerer and Hamill, 2015). For an ensemble forecast $\mathbf{f}_m = (f_m^{(1)}, f_m^{(2)}, \dots, f_m^{(D)})^\top$, $m = 1, 2, \dots, M$, it is defined as

$$\text{VS}^p(F_M, \mathbf{y}) = \sum_{i=1}^D \sum_{j=1}^D \omega_{ij} \left(|y^{(i)} - y^{(j)}|^p - \frac{1}{M} \sum_{m=1}^M |f_m^{(i)} - f_m^{(j)}|^p \right)^2, \quad (3.1.6)$$

where $\omega_{ij} \geq 0$ is the weight for coordinate pair (i, j) . Compared to the ES, the VS^p is more sensitive to the errors in the specification of correlations.

Further, we also investigate the mean Euclidean distance (EE) of the median vectors of the forecasts from the corresponding validating observations, where the multivariate L^1 ensemble median can be obtained e.g. with the algorithm of Vardi and Zhang (2000).

Multivariate generalizations of the verification rank are based on the concept of pre-ranks and the corresponding histograms display the pre-rank of the observation with respect to the ensemble forecast with ties resolved at random (Gneiting *et al.*, 2008). Here we consider the average and band-depth ranks suggested by Thorarinsdottir *et al.* (2016) and energy score and dependence

(variogram) ranks studied in detail e.g. by Allen *et al.* (2024). The average rank is simply the mean of the univariate ranks and the interpretation of the corresponding histogram is similar to the univariate case: $\cup$ - and $\cap$ -shaped histograms refer to under- and overdispersion, respectively, whereas biased forecasts result in triangular shapes. Band-depth histograms are based on the discrete version of ordering multivariate functional data proposed by López-Pintado and Romo (2009), while energy score and dependence histograms use the idea of Knüppel *et al.* (2022) of proper score-based pre-ranks. Right-skewed band-depth histograms indicate overdispersion, left-skewed ones underdispersion or bias, whereas $\cup$ - and $\cap$ -shaped histograms mean too low- and high correlations in the ensemble. Furthermore, low score-based ranks indicate that the observation is similar to the ensemble members in terms of the given scoring rule, whereas outliers result in high pre-ranks. Here, we would like to highlight that reliability indices can also serve as an effective tool for ranking multivariate predictions. However, the interpretation of the resulting rank histograms may vary, as the functions employed to calculate the pre-ranks can differ in their sensitivity to specific deficiencies, potentially impacting the conclusions drawn.

For a discrete forecast F represented by a PMF $p_F(y)$ on $\mathcal{Y} = \{y_1, y_2, \dots, y_C\}$ – that defines the observable values – the logarithmic score and the continuous ranked probability score is defined as

$$\text{LogS}(F, y) := -\log(p_F(y)), \quad (3.1.7)$$

whereas the latter equals

$$\text{CRPS}(F, y) = \sum_{k=1}^C p_F(y_k) |y - y_k| - \sum_{k=2}^C \sum_{\ell=1}^{k-1} p_F(y_k) p_F(y_\ell) |y_\ell - y_k|, \quad (3.1.8)$$

which is the discrete version of the general representation (3.1.1) derived by Gneiting and Raftery (2007).

Further, in our case studies, for a given forecast F the improvement in terms of a score $\mathcal{S}_F$ with respect to a reference forecast F_{ref} is quantified using the corresponding skill score (Gneiting and Raftery, 2007)

$$\mathcal{S}_F^{\text{skill}} := 1 - \frac{\overline{\mathcal{S}}_F}{\overline{\mathcal{S}}_{F_{\text{ref}}}},$$

where $\overline{\mathcal{S}}_F$ and $\overline{\mathcal{S}}_{F_{\text{ref}}}$ denote the mean score values over all forecast cases in the verification period for forecasts F and F_{ref} , respectively. Thus, besides the mean scores, one can investigate the corresponding skill scores as well, that are positively oriented (the larger the better).

Finally, following Gneiting and Ranjan (2011), statistical significance of score differences between forecasts are assessed with the help of the Diebold-Mariano (DM; Diebold and Mariano, 1995) test, which is able to account for temporal dependencies in the forecast errors. Given a scoring rule $\mathcal{S}$ and two competing probabilistic forecasts F and G , the test statistic of the DM test is given by

$$t_N = \sqrt{N} \frac{\bar{\mathcal{S}}_F - \bar{\mathcal{S}}_G}{\hat{\sigma}_N}, \quad (3.1.9)$$

where $\bar{\mathcal{S}}_F$ and $\bar{\mathcal{S}}_G$ are the mean scores over a test set corresponding to forecasts F and G , respectively, and $\hat{\sigma}_N$ is a suitable estimator of the asymptotic standard deviation of the sequence of individual score differences. Under standard regularity conditions, t_N asymptotically follows a standard Gaussian distribution under the null hypothesis of equal predictive performance. Negative values of t_N indicate a better predictive performance of F , whereas G is preferred in the case of positive values of t_N . Following the suggestions of Wilks (2016), in simultaneous testing for the various locations at a 5 % level of significance, the false discovery rate can be controlled by the Benjamini-Hochberg algorithm (Benjamini and Hochberg, 1995).

Another method for examining statistical significance is to report block bootstrap confidence intervals for the skill scores and score differences. These intervals are based on numerous samples calculated using the stationary bootstrap scheme with an average block length determined according to Politis and Romano (1994).

3.2 Training data selection

Training data selection is a fundamental aspect of building effective statistical models, just as it is in machine learning. In the case of ensemble calibration methods (e.g. EMOS, POLR or MLP), selecting the right training data is essential for achieving precise calibration, effective model fitting, and dependable probabilistic forecasts. Using high-quality training data allows these methods to accurately align ensemble forecasts with observed outcomes and manage a variety of data distributions. The selection process consists of two main steps, one of which is determining the spatial selection strategy. The typical options include three alternatives: local, regional, and semi-local method that falls between the two. Regional modelling (Thorarinsdottir and Gneiting, 2010) involves calibrating forecasts based on the entire ensemble domain. This approach implies that when calibrating forecasts for a particular observation station, data from all other available stations within the domain are utilized to estimate the parameters, resulting in the same parameters across all locations. While regional models are designed to cover the entire available

geographic area to offer detailed and relevant insights, this broad focus can sometimes constrain the model’s flexibility compared to more localized modelling approaches. In contrast to regional modelling, local modelling focuses on the historical data from a specific observation station. While this approach involves significantly fewer training data, it often results in more accurately calibrated predictive distributions. In general, locally trained models result in the best forecast skill, provided the training period is long enough to circumvent numerical problems in modelling. However, in areas with high variability, local modelling can be ineffective. To combine the beneficial properties of local and regional modelling, Lerch and Baran (2017) propose a clustering-based semi-local approach. Based on individual feature vectors assigned to each observation station, which represent climatology and performance of the ensemble forecast during the training period, locations are first arranged into clusters using k -means clustering (see e.g. Wilks, 2019, Section 16.3.1). Then, within each cluster, a regional estimation is performed resulting in a separate set of parameters for each cluster. Another challenge pertains to the temporal selection of training data. Standard practice for estimating EMOS model parameters involves using a rolling window approach. In this method, estimates are derived from ensemble forecasts and corresponding validation observations from the previous n calendar days. Specifically, for a given verification day, the training data consists of the n days preceding it, and this window shifts forward one day with each new verification day. This popular approach allows quick adaptation, for instance, to model changes or seasonal variations. However, one can also consider monthly, seasonal, or yearly windows (see e.g. Hemri *et al.*, 2016; Jobst *et al.*, 2023). On the other hand, machine learning-based techniques, which require large training datasets, usually rely on static training data (see e.g. Schultz and Lerch, 2022).

Chapter 4

Clustering-based spatial interpolation of parametric post-processing models

Post-processing techniques, like EMOS, usually depend on training data made up of past forecast-observation pairs, meaning calibrated forecasts are typically generated only for locations where such data is available. Although forecasts are produced at grid points with a specific resolution and can be bilinearly interpolated to any location, surface observations are mainly collected at synoptic observation (SYNOP) stations, which are unevenly scattered. While gridded reanalyses, which align with the spatio-temporal scale of forecasts, can serve as verification data, station observations are generally preferred for training and validation, as they result in more precise post-processing.

This chapter aims to investigate how station-based post-processing, specifically EMOS modelling of wind speed ensemble forecasts, can be extended to locations lacking observation data. Our method can be easily adapted to other parametric approaches and weather variables, without requiring modifications to the interpolation techniques introduced in Section 4.1, offering a more straightforward alternative to kriging-based methods, such as those proposed by Scheuerer and Büermann (2014) and Scheuerer and Möller (2015), which involve complex Gaussian random fields and additional covariates for interpolation. Additionally, it contrasts with the local BMA method of Kleiber *et al.* (2011), which merges Mass-Baars interpolation with a regionally estimated BMA model. In a related study (Baran and Lakatos, 2024), we focused on the truncated normal (TN) EMOS model (see Section 1.2.2), which estimates distribution parameters based on historical forecast and observation data. We illustrate the effectiveness of our approach through a case study using 10-meter

wind speed ensemble forecasts from the ECMWF. Our findings reveal that this method surpasses both regionally estimated (discussed in Section 3.2) and interpolated EMOS models, as well as raw ensemble forecasts.

In the regional case, the common set of parameters obtained can be utilized to determine the EMOS predictive distribution at any location where forecasts are provided. Prior research has demonstrated (Baran and Baran, 2024) that regional post-processing and subsequent interpolation can greatly enhance forecast accuracy, with improvements exceeding 20% at observed locations and over 11% at unobserved sites.

However, in large ensemble domains, differences in climatology and forecast error characteristics among various observation sites can make a single set of EMOS parameters from regional approaches less effective. In such cases, relying solely on local data for parameter estimation at the specific station is a practical alternative. However, directly interpolating these local EMOS models to locations without observations presents a challenge. To address this, we propose an innovative semi-local modelling method that enhances the clustering-based semi-local approach of Lerch and Baran (2017) by adding an interpolation component. This technique combines the strengths of both regional and local parameter estimation. At observed locations, the method uses k -means clustering to form groups of similar sites based on training data features, with each cluster sharing the same set of EMOS parameters. The key innovation is the interpolation process, where parameters for unobserved sites are assigned based on the nearest cluster, determined solely by training forecast data.

We compare the predictive performance of our semi-local method with regional and local EMOS models, as well as with raw forecasts. The evaluation involves developing EMOS models using data from 75% of SYNOP stations and then interpolating these models to the remaining 25% of sites.

4.1 Interpolation algorithms

As detailed in Section 3.2, the parameters for all EMOS models, including the truncated normal model, are estimated using training data, where both temporal and spatial factors can affect the accuracy of post-processed forecasts. In our study, we employ rolling training periods based on ensemble forecasts and corresponding observations from the preceding n calendar days.

The feature selection method described by Lerch and Baran (2017) in Section 3.2, referred to as *observation-based* and discussed further below, has demonstrated effectiveness in various case studies involving different weather variables, such as temperature (Baran *et al.*, 2019), heat indices (Baran *et al.*, 2020), and precipitation (Szabó *et al.*, 2023).

Observation-based quantile features

Following the notations of Lerch and Baran (2017), let $\hat{F}_{i,n}$ and $\hat{G}_{i,n}^e$ denote the empirical CDFs of station climatology (observed 10-m wind speed) and forecast errors of the mean $\bar{f}_{\text{ENS}}$ of the exchangeable ensemble members at location i over the rolling training period of length n , respectively. The N -dimensional feature vector (referred to as OQ) assigned to station i is defined as the concatenation of equidistant $\frac{1}{N_1+1}, \frac{2}{N_1+1}, \dots, \frac{N_1}{N_1+1}$ quantiles of $\hat{F}_{i,n}$ and equidistant $\frac{1}{N_2+1}, \frac{2}{N_2+1}, \dots, \frac{N_2}{N_2+1}$ quantiles of $\hat{G}_{i,n}^e$, where $N_1 = \lfloor \frac{N}{2} \rfloor$ and $N_2 = N - N_1$.

From the perspective of model interpolation, the disadvantage of the above feature set is that it uses observations, hence it is not straightforward how to assign an unobserved location to one of the existing clusters of observed locations, where model parameters are available. A possible solution is to consider features that are based solely on the ensemble forecasts. This approach will be called *forecast-based* and we consider two different setups.

Forecast-based quantile features

This feature set (referred to as FQ) is the forecast-based analogue of the previous, observation-based construction. Denote by $\hat{H}_{i,n}^{(1)}$, $\hat{H}_{i,n}^{(2)}$, $\hat{H}_{i,n}^{(3)}$ and $\hat{H}_{i,n}^{(4)}$ the empirical CDFs of high-resolution forecasts f_{HRES} , control forecasts f_{CTRL} and mean $\bar{f}_{\text{ENS}}$ and standard deviation S_{ENS} of the 50 exchangeable ensemble members, respectively, at location i over the training period of length n . Now the N dimensional feature vector for station i comprises the concatenation of equidistant quantiles of $\hat{H}_{i,n}^{(j)}$ at levels $\frac{1}{N_j+1}, \frac{2}{N_j+1}, \dots, \frac{N_j}{N_j+1}$, $j = 1, 2, 3, 4$, where $N_1 = N_2 = N_3 = \lfloor \frac{N}{4} \rfloor$ and $N_4 = N - N_1 - N_2 - N_3$.

Forecasts as features

Instead of equidistant quantiles, for each location i and training period of length n , consider simply the $4n$ -dimensional concatenation of training vectors of f_{HRES} , f_{CTRL} , $\bar{f}_{\text{ENS}}$ and S_{ENS} (the n high-resolution and n control forecasts from the training window and the n -dimensional mean and standard deviation vectors of the corresponding exchangeable members) in the order as they are listed. Naturally, this simple approach might result in rather high-dimensional features (referred to as FF), which increases the computational cost of k -means clustering. However, the additional cost is usually less than the cost of computing the equidistant quantiles comprising a forecast-based quantile feature.

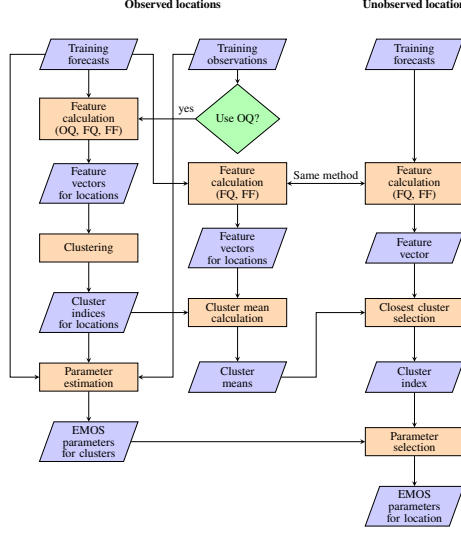


Figure 4.1: Flowchart of the semi-local parameter estimation method for the observed locations and the interpolation algorithm for an unobserved site.

Interpolation algorithm

As forecast-based features are available at both observed and unobserved locations, we propose the following interpolation algorithm, an overview of which is provided in the flowchart of Figure 4.1.

Step 1. Using either observation-based features or one of the forecast-based approaches, estimate clustering-based semi-local EMOS models for the observed locations. As outputs, one gets the clusters of observed sites and the parameters of the corresponding EMOS models, see the left column of Figure 4.1.

Step 2. Choose one of the forecast-based feature selection methods (not necessarily the same as in Step 1). To each observed location assign the corresponding feature vector and calculate the cluster means for the clusters formed in Step 1. The outputs of this step are these cluster means, see the middle column of Figure 4.1.

Step 3. Calculate the same forecast-based features as in Step 2 for unobserved locations (operation *feature calculation* in the right column of Figure 4.1).

Step 4. Assign each unobserved location to the closest cluster in terms of the Euclidean distance of the feature vector of the given unobserved location to the cluster mean (operation *closest cluster selection* in the right column of Figure 4.1).

Step 5. For each unobserved location use the EMOS parameters of the assigned cluster to calculate the corresponding predictive distribution, see the final output of the the right column of Figure 4.1.

When the number of clusters matches the number of observed locations, the approach reverts to a local method. Conversely, using a single cluster aligns with regional estimation. In the local method, unobserved locations are assigned to the nearest observed one based on the Euclidean distance between their feature vectors.

An alternative would be to use geographical distance for interpolation. This method assigns an unobserved location to the closest cluster based on the average geographical distance to cluster members. For local models, this would mean that an unobserved location inherits the EMOS parameters of the geographically nearest observed location. However, for our dataset, geographical interpolation consistently falls short compared to the forecast-based algorithm across all semi-local cases and lead times, based on three case studies discussed in Section 4.3. Although geographical matching offers slight benefits for local models at short lead times, this minor advantage is still inferior to forecast-based matching that minimizes the Euclidean distance between the corresponding FQ or FF feature vectors, and overall, local models do not match the performance of interpolated forecast-based semi-local methods. Additionally, our study found no evidence of a negative correlation between the predictive accuracy of a calibrated forecast at an unobserved location and its geographical distance from the nearest observed site providing the EMOS parameters.

4.2 Data

In our case study, we assess the efficacy of the interpolation algorithms delineated in Section 4.1 utilizing 52-member ECMWF ensemble forecasts of 10-meter wind speed for 674 SYNOP stations across Europe, Asia, and Australia, covering the period from January 1, 2020, to April 30, 2023. This dataset includes the ECMWF HRES forecast at TCo1279 resolution (9 km grid), the CNTL forecast, and 50 ENS members generated with perturbed initial conditions, all with an 18-km grid spacing (TCo639). Forecasts are initialized at 0000 UTC, and we examine 10 different lead times ranging from 1 to 10 days. Out of the total SYNOP stations, 75% (506 stations, referred to as "*observed locations*") were randomly selected to estimate the parameters for wind speed predictive distributions. These models were then interpolated to the remaining

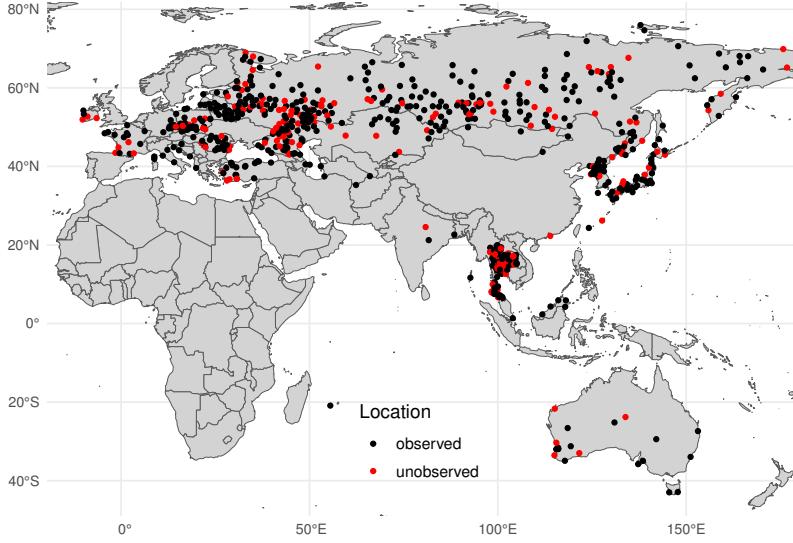


Figure 4.2: Locations of SYNOP stations and the distribution of observed and unobserved locations.

168 stations (referred to as "*unobserved locations*"), as shown in Figure 4.2. The dataset has about 1.65% of the forecast cases incomplete due to missing forecasts or observations.

4.3 Results

The quality of the interpolation algorithm presented in Section 4.1 in terms of the predictive performance of post-processed forecasts at unobserved locations is tested with the help of ECMWF 10-m wind speed data introduced in Section 4.2. Based on the three investigated feature sets, we consider three different semi-local EMOS models for the observed locations (Step 1), which will be referred to as *semi-local OQ* (observation-based quantile features), *semi-local FQ* (forecast-based quantile features) and *semi-local FF* (forecasts as features). According to Lerch and Baran (2017), provided one has enough features to represent the empirical CDFs, the size N of the quantile-based features has only a marginal effect. Hence, in the observation-based case we set $N = 24$ as in Lerch and Baran (2017) and Szabó *et al.* (2023), whereas in the forecast-based one, similar to Baran *et al.* (2020), 40-dimensional features are considered. Observed locations are arranged into 20 clusters, and unobserved locations are assigned to clusters (Steps 2 and 3) utilizing forecasts as features, since this approach results in more skillful models than the use of forecast-based quantile

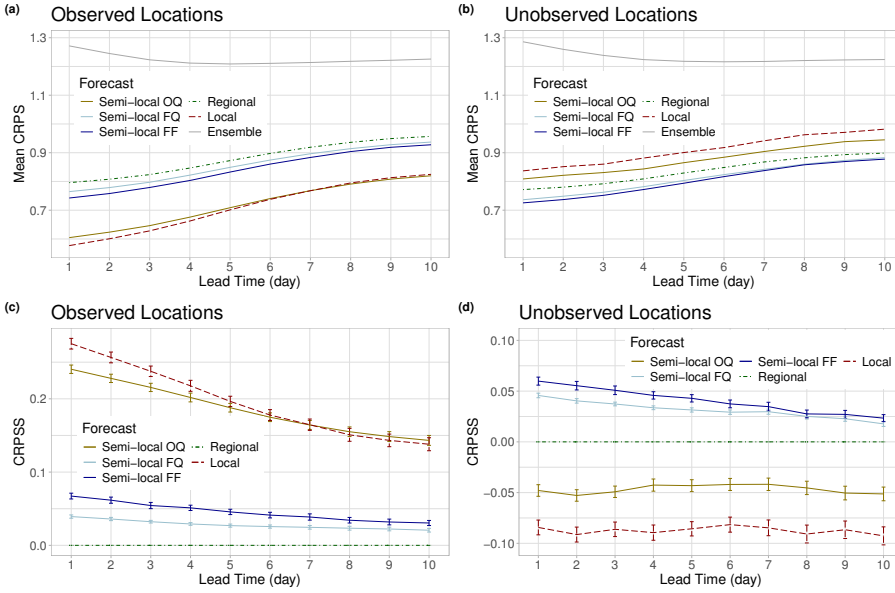


Figure 4.3: CRPS of post-processed and raw 10-m wind speed forecasts for observed (a) and unobserved (b) locations, and CRPSS of local and semi-local models with respect to the regional approach for observed (c) and unobserved (d) locations together with 95% confidence intervals as functions of the lead time.

features. As a reference post-processing approach, we consider the regional TN EMOS model based only on the observed locations and interpolated to the unobserved sites. As mentioned, in the regional case one has a single set of EMOS parameters for all observed sites, which parameters are then utilized at the unobserved locations as well. Furthermore, we also report the scores of the local TN EMOS model and its interpolation with the help of forecasts as features (an unobserved location inherits the parameters of the closest observed one in terms of the Euclidean distance of their feature vectors, see Section 4.1), and the raw ECMWF ensemble forecasts. All EMOS models are trained using a 60-day rolling training window, and forecast-observation pairs of the period May 1, 2020 – April 30, 2023 are considered for validation purposes. Note that semi-local models based on 10 and 15 clusters and training period lengths of 30 and 90 days were also tested. In general, the predictive performance of the EMOS models exhibits only a very slight dependence on the training period length with 60 days being the optimal one, and fewer clusters result in slightly worse forecast skill for all investigated forecast-based semi-local models and

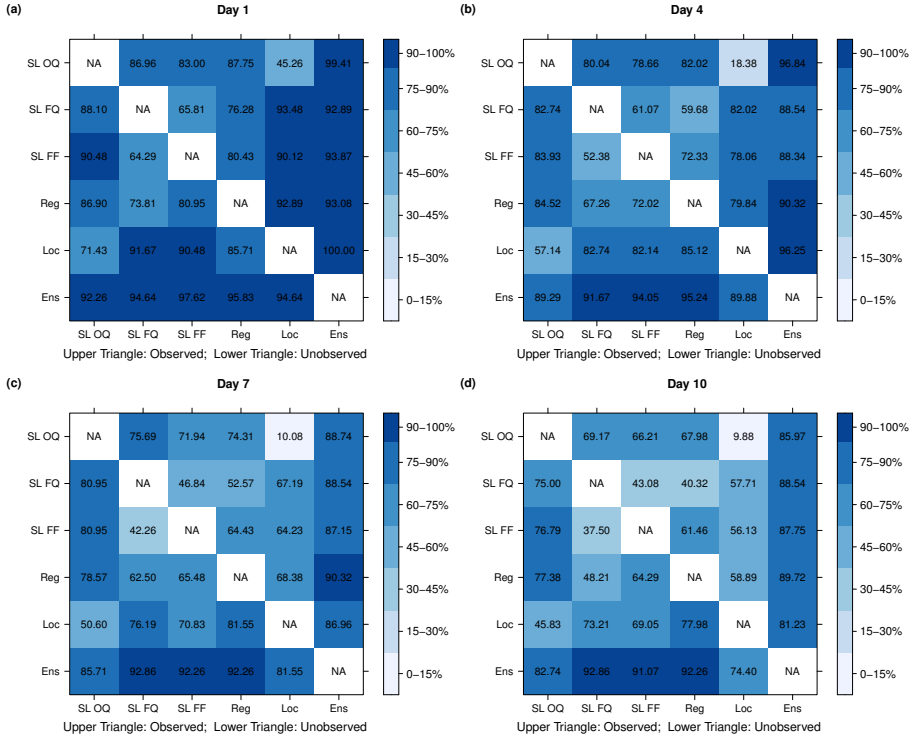


Figure 4.4: Proportions of observed (upper triangle) and unobserved (lower triangle) locations with significantly different mean CRPS at a 5% level of significance for lead times 1 (a), 4 (b), 7 (c) and 10 (d) days.

training windows.

Figures 4.3a,b show the mean CRPS of post-processed and raw forecasts for observed and unobserved locations, respectively. As expected, all calibrated forecasts outperform the raw ensemble by a wide margin, and the skill of the various EMOS models degrades with the increase of the forecast horizon. Note that the different shape of the mean CRPS of the raw forecasts is due to the representativeness error in the verification, which behaviour has been observed in other case studies as well (see e.g. Baran *et al.*, 2021), and can be partially corrected by the perturbed-ensemble approach of Ben Bouallègue *et al.* (2020). At the observed locations the local TN EMOS model exhibits the best overall predictive performance, closely followed by the observation-based semi-local approach (OQ) of Lerch and Baran (2017), which after day 7 results in the lowest mean CRPS. Furthermore, according to the skill scores of Figure 4.3c,

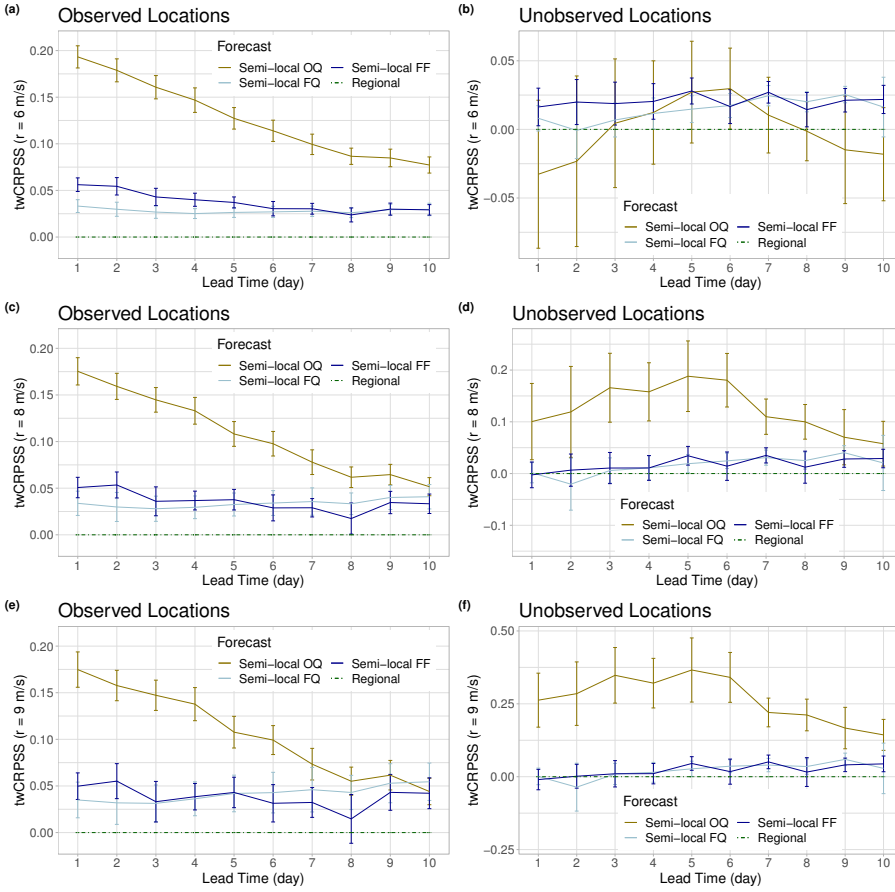


Figure 4.5: twCRPSS values of semi-local models with respect to the regional approach for thresholds 6 m/s (a,b), 8 m/s (c,d) and 9 m/s (e,f) for observed (a,c,e) and unobserved (b,d,f) locations together with 95% confidence intervals as functions of the lead time.

at observed locations all semi-local EMOS models significantly outperform the regional one with the observation-based (OQ) approach being far the best, and the difference between the two forecast-based methods is also significant for all lead times. As Figures 4.3b,d indicate, the lack of observations drastically changes the model ranking. Local model results in the highest mean CRPS and while the performance of the two forecast-based semi-local EMOS models is similar at observed and unobserved locations, the observation-based semi-local EMOS is significantly behind its competitors. In general, at unobserved

locations the skill of best performing forecast-based semi-local FF EMOS approach is 2.3 – 6 % ahead of the performance of the regional TN EMOS model, whereas at observed sites its advantage is in the 3 – 6.7 % interval.

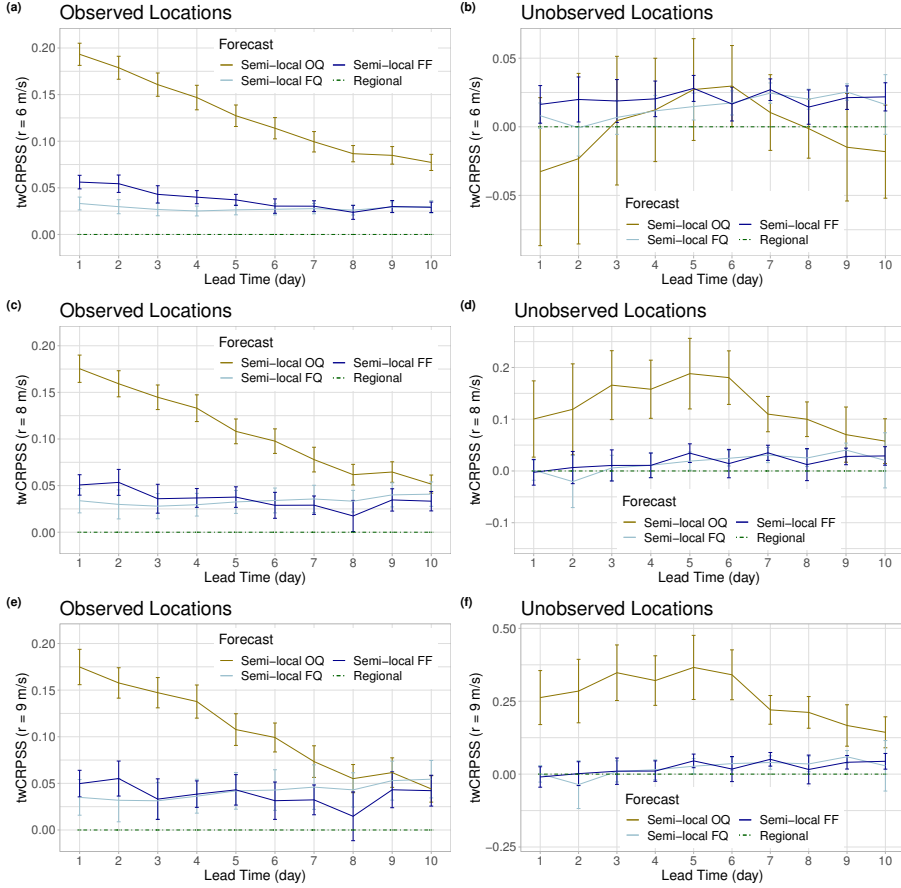


Figure 4.6: twCRPSS values of semi-local models with respect to the regional approach for thresholds 6 m/s (a,b), 8 m/s (c,d) and 9 m/s (e,f) for observed (a,c,e) and unobserved (b,d,f) locations together with 95 % confidence intervals as functions of the lead time.

The above findings are further supported by Figure 4.4 showing the proportions of stations where simultaneous DM tests for observed and unobserved locations indicate significant difference in the mean CRPS over the verification period at a 5 % level for lead times 1, 4, 7 and 10 days. The raw ensemble is clearly separated from all other forecasts for all lead times, whereas the de-

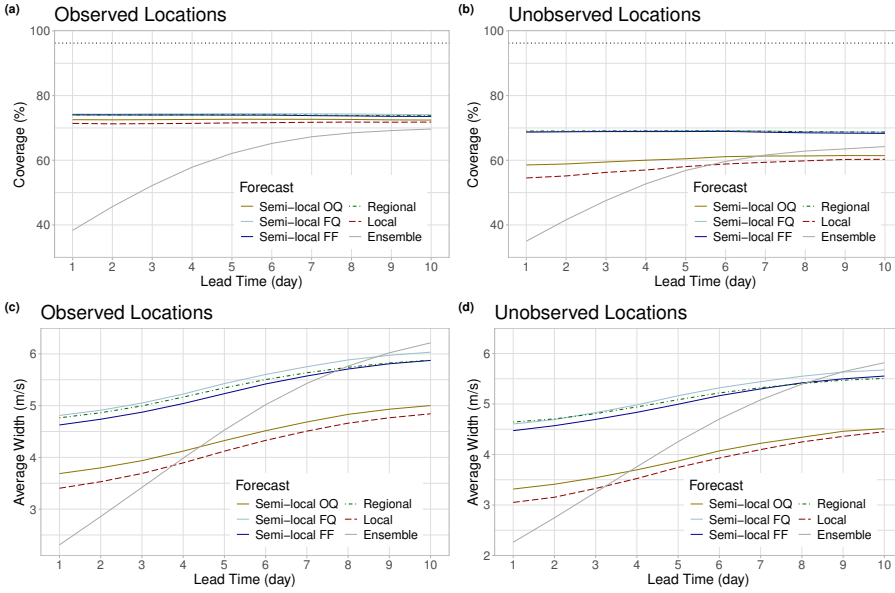


Figure 4.7: Coverage (a,b) and average width (c,d) of nominal 96.23 % central prediction intervals of post-processed and raw 10-m wind speed forecasts for observed (a,c) and unobserved (b,d) locations functions of the lead time. In panels (a) and (b) the ideal coverage is indicated by the horizontal dotted line.

ing proportions of observed locations where local and semi-local OQ forecasts differ significantly (day 1: 45.26 %; day 4: 18.38 %; day 7: 10.08 %; day 10: 9.88 %) are nicely in line with the behaviour of the corresponding curves of Figures 4.3a,c.

To compare the skill of post-processed forecasts in predicting high wind speeds, we consider the twCRPSS values of the semi-local methods with respect to the regional EMOS approach for thresholds 6, 8 and 9 m/s corresponding to the 95th, 98th, and 99th percentiles of observed 10-m wind speed, respectively (Figure 4.6). Here the local model is omitted from the analysis as at unobserved locations in terms of the mean twCRPS it is far behind the competitors for all thresholds. At observed locations, the ranking of the semi-local models is rather consistent (Figures 4.6a,c,e) and resembles the ordering with respect to the CRPSS (see Figure 4.3c). The observation-based (OQ) semi-local modelling is superior to the forecast-based (FQ and FF) approaches, which are quite close to each other in forecast skill. However, the forecast uncertainty increases with the threshold and for $r = 9$ m/s from day 8 the difference between the semi-local OQ and semi-local FQ forecasts is not significant. In contrast, at

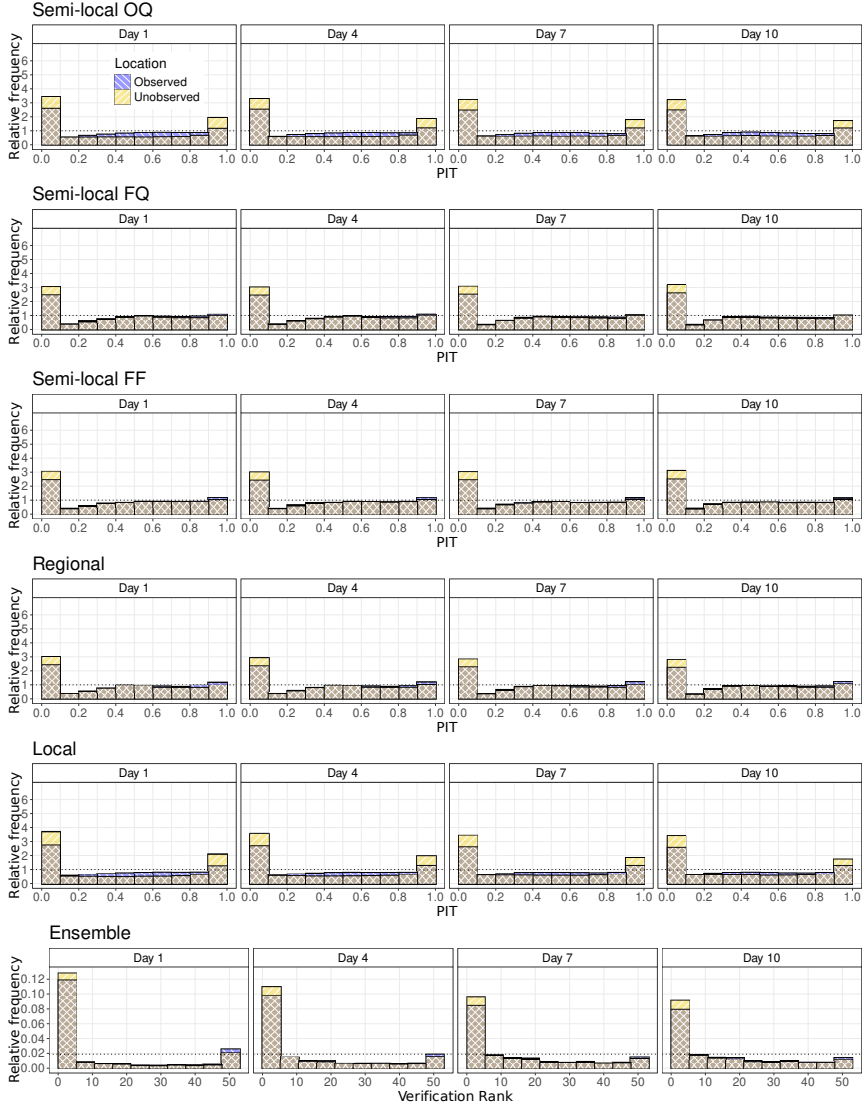


Figure 4.8: PIT histograms of post-processed and verification rank histograms of raw ensemble forecasts of 10-m wind speed for lead times 1, 4, 7 and 10 days. The ideal uniform distribution is indicated by the horizontal dotted line. The cross-hatched parts indicate the intersection of the corresponding histograms for observed and unobserved locations.

Forecast	Observed locations				Unobserved locations			
	Day 1	Day 4	Day 7	Day 10	Day 1	Day 4	Day 7	Day 10
Semi-local OQ	0.117	0.116	0.115	0.115	0.148	0.145	0.140	0.138
Semi-local FQ	0.112	0.112	0.112	0.112	0.117	0.116	0.117	0.118
Semi-local FF	0.114	0.114	0.114	0.114	0.119	0.119	0.119	0.120
Regional	0.113	0.112	0.111	0.111	0.115	0.116	0.116	0.116
Local	0.121	0.121	0.119	0.119	0.156	0.151	0.145	0.141

Table 4.1: Variances of PIT values of post-processed forecasts of 10-m wind speed for lead times 1, 4, 7 and 10 days. The reference variance of the ideal standard uniform distribution is 0.083.

unobserved locations the predictive performance of the semi-local forecasts at the tails (Figures 4.6b,d,f) completely differs from their overall performance (Figure 4.3d). While in terms of the mean CRPS the semi-local OQ forecast is significantly behind the regional and forecast-based semi-local (FQ and FF) methods for all lead times, for the lowest threshold of 6 m/s the differences in twCRPSS are not significant, whereas for higher thresholds observation-based (OQ) semi-local modelling results in significantly higher twCRPSS values than the competing EMOS models.

As mentioned in Chapter 3, the general aim is to obtain predictive distributions that are calibrated, meaning around 96.23% of verifying observations fall in the nominal 96.23% symmetric central prediction intervals that are, on average, narrower than those of raw ensemble forecasts. As displayed in Figure 4.7a, at observed locations all post-processing approaches substantially improve the coverage of the raw ECMWF 10-m wind speed forecasts, which ranges from 38.32 % to 69.65 % and increases with the increase of the forecast horizon. Regional and forecast-based semi-local EMOS models (FQ and FF) result in the highest and almost identical coverage values with a slightly decreasing trend; however, their coverage of around 74 % is still far from the nominal 96.23 % level. Note that the coverages of observation-based semi-local (OQ) and local EMOS approaches – resulting in far the sharpest predictive distributions (see Figure 4.7c) – are just slightly below the coverage values of the other three EMOS models. According to Figures 4.7b,d, at unobserved locations the behaviour of raw forecasts and interpolated regional and forecast-based semi-local EMOS models are rather similar to what we see at the observed sites (Figures 4.7a,c); nonetheless, there is a slight systematic decrease of around 4 percentage points in all coverage values. In contrast, the coverages of the local and observation-based semi-local EMOS models are substantially lower; from days 6 and 7, respectively, they are even below the coverage of the raw ensemble. This means, that for longer lead times at unobserved locations these latter two post-processing methods can hardly improve the statistical inconsistency of the

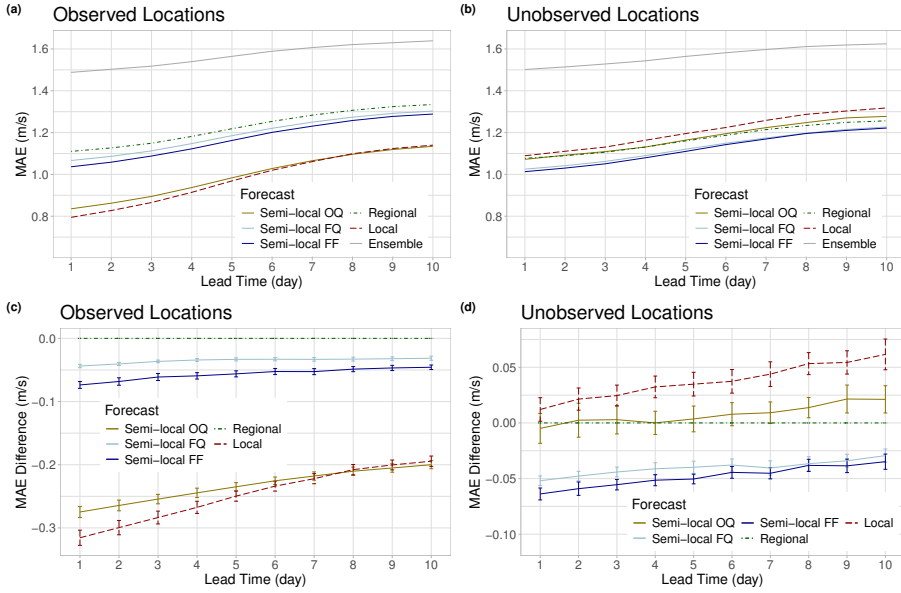


Figure 4.9: MAE of the median 10-m wind speed forecasts for observed (a) and unobserved (b) locations, and difference in MAE of local and semi-local models from the regional approach for observed (c) and unobserved (d) locations together with 95 % confidence intervals as functions of the lead time.

raw ECMWF predictions and the corresponding observations.

The verification rank and PIT histograms of Figure 4.8 provide further evidence of the improved calibration of post-processed forecasts compared to the raw ECMWF ensemble both at observed and unobserved locations. Raw forecasts are highly underdispersive (U-shaped) with a strong positive bias at both sets of sites; nevertheless, these deficiencies become less pronounced with the increase of the forecast horizon. Post-processing considerably decreases underdispersion and to a certain extent also the bias. Regional and forecast-based semi-local methods result in rather similar PIT histograms both at observed and unobserved locations; however, in the latter case, the bias correction is slightly less successful. In contrast, at unobserved locations, the PIT histograms of local and observation-based semi-local (OQ) EMOS models are far more underdispersive than their counterparts at the observed sites. The variances of the PIT values reported in Table 4.1 are nicely in line with the above findings. All displayed variances are much larger than the reference variance of 0.083 (1/12) of the standard uniform distribution, indicating that the corresponding predictive distributions are underdispersive (see e.g. Möller and

Groß, 2020). The smallest PIT variances correspond to the regional and semi-local FQ models both at observed and unobserved sites, closely followed by the semi-local FF approach.

Finally, Figures 4.9a,c showing the MAE of median forecasts and the difference in MAE of semi-local EMOS approaches from the regional EMOS tell the same story as the corresponding panels of Figure 4.3. At observed locations local and observation-based semi-local EMOS calibration is the most beneficial and the MAE values of regional and semi-local models differ significantly. However, in contrast to the mean CRPS (Figures 4.3b,d), while the behaviour and ranking of regional and forecast-based semi-local models remain the same at observed and unobserved sites (Figures 4.9b,d), at unobserved locations up to day 7, there is no significant difference in terms of MAE between the regional and the observation-based semi-local EMOS model.

Based on the above analysis one can conclude that while at observed locations the local approach exhibits the best overall performance, closely followed by the observation-based semi-local modelling, clustering-based interpolation of semi-local EMOS models considering forecasts as features is superior to the regional approach to parameter estimation and results in more skillful predictive distributions at locations where no observations are available.

Finally, note that we also experimented with the kriging-based technique of Scheuerer and Möller (2015) with the intention of using it as a further reference method. At observed locations we considered the suggested two-step EMOS approach with a 60-day rolling training period, whereas at unobserved sites the location and the log-scale of the TN predictive distribution were modelled by intrinsically stationary fractional Brownian surfaces with nugget effect. In general, at observed locations the two-step TN EMOS model of Scheuerer and Möller (2015) exhibits a fair predictive performance ranked above our regional and forecast-based semi-local TN EMOS methods, but below the local and observation-based semi-local TN EMOS forecasts. In contrast, at unobserved sites the kriging-based prediction substantially underperforms all competitors; moreover, the estimation of the covariance parameters of the underlying random fields and the kriging processes require several magnitudes higher computation time than our interpolation algorithm. For instance, for a single lead time, the kriging-based interpolation for all 1095 dates in the verification period takes approximately 6.48 hours, whereas our approach requires roughly 2 minutes (R version 4.1.2; Ubuntu Linux 22.04.4 LTS, Intel(R) Core(TM) i7-9700KF CPU, 3.60GHz; 64 Gb RAM). Hence, for the sake of brevity, we decided against providing the details of the kriging-based interpolation, and reporting the corresponding verification results.

4.4 Summary and conclusion

In this chapter, we presented a new methodology for interpolating parametric predictive distributions that avoids the computational cost of the spatial interpolation methods proposed either by Kleiber *et al.* (2011) or by Scheuerer and Büermann (2014) and Scheuerer and Möller (2015), since the presented methodology does not need the computationally very expensive estimation of an underlying Gaussian field or require any additional station-specific information. This approach, which relies solely on the data essential for EMOS modelling, is adaptable to any parametric post-processing method, including distributional regression networks (Rasp and Lerch, 2018).

Our results indicate that local and observation-based semi-local methods perform best at observed sites, while forecast-based semi-local approaches outperform both observation-based and local methods at unobserved locations. Compared to regional models, forecast-based semi-local models offer sharper prediction intervals, better mean CRPS, and lower MAE of median forecasts.

These findings highlight that no universal post-processing method is ideal for all scenarios. For observed sites, local or observation-based semi-local methods are preferable, while forecast-based semi-local methods are more effective for unobserved locations. The success of forecast-based methods is due to the alignment of feature vectors derived from similar quantities.

Future research could explore extending station-based post-processing to gridded dual-resolution forecasts, especially with the new CY48R1 ECMWF model cycle, and examine the performance of multivariate post-processing methods, such as ensemble copula coupling or Gaussian copula approaches, for unobserved and gridded locations.

Chapter 5

Statistical post-processing of visibility ensemble forecasts

5.1 Introduction

Visibility is a critical parameter in aviation meteorology, significantly impacting all phases of flight, particularly during takeoff and landing. Low visibility conditions necessitate the implementation of Low Visibility Procedures (LVPs) (see e.g. Roquelaure and Bergot, 2009; Boudala *et al.*, 2012; Dietz *et al.*, 2019) at airports, which can reduce capacity and potentially lead to delays or cancellations. Beyond aviation, visibility plays a vital role in maritime navigation, road transportation, and agriculture, with poor visibility being one of the leading causes of road accidents.

According to the World Meteorological Organization (WMO), visibility is defined as “the greatest distance at which a black object of suitable dimensions, located on the ground, can be seen and recognized when observed against the horizon sky” (WMO, 1992). Therefore, accurate visibility forecasts yield direct economic and safety benefits across various sectors. Despite advancements in meteorological forecasting, visibility remains a particularly challenging variable to forecast accurately. Most NWP models do not explicitly simulate visibility, instead deriving forecasts from other weather variables such as relative humidity, precipitation, or aerosol concentration (Chmielecki and Raftery, 2011). Several factors contribute to the complexity of forecasting visibility, including challenges in simulating fog formation and resolving microphysical processes that affect visibility in polluted environments (Gultepe *et al.*, 2006; Wagh *et*

al., 2023). Additionally, the coarse vertical resolution of operational NWP models can hinder the accurate representation of fog formation processes. At the ECMWF, visibility was incorporated into the Integrated Forecasting System (IFS) in 2015, based on model outputs of water vapor, clouds, precipitation, and climatological aerosol fields (ECMWF, 2024a). Visibility observations, particularly at SYNOP stations, are typically reported in discrete values as recommended by the WMO: “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” (WMO, 2018, Section 9.1.2). A much finer, but still discrete reporting scheme is used in the Meteorological Aerodrome Report (METAR), where the maximal reported visibility is 10 km, 50 m steps are considered up to 800 m, 100 m steps from 800 m to 5 km, and 1 km steps from 5 km to 10 km (WMO, 2022). Hence, similarly to total cloud cover (TCC) forecasts, which are reported in oktas, visibility post-processing can be approached as a classification task, estimating probabilities for discrete categories. This framework has been effectively utilized in post-processing TCC forecasts using methods such as POLR and MLP neural networks (see Section 1.3), both of which have demonstrated superior forecast skill compared to other methods (Baran *et al.*, 2021).

In this chapter, we focus on calibrating visibility ensemble forecasts from the ECMWF and utilizing the EUPPBench dataset (Demaeyer *et al.*, 2023), employing POLR and MLP approaches (Baran and Lakatos, 2023). We explore the potential benefits of incorporating visibility predictions from the Copernicus Atmosphere Monitoring Service (CAMS) as additional covariates to enhance overall forecast accuracy (including low visibility events). Throughout our analysis, we use raw ensemble forecasts and climatology as reference benchmarks. Additionally, we investigate whether the inclusion of CAMS data helps reinstall spatial dependencies between marginal distributions, thereby addressing a key limitation of univariate post-processing methods (Lakatos and Baran, 2024).

5.2 Data

In the subsequent subsections, we examine two datasets containing ECMWF visibility ensemble forecasts and their corresponding validation observations, which we calibrate using the classification methods outlined in Section 1.3. Although these datasets encompass different time periods, there are some overlapping areas within their ensemble domains. The observations follow WMO guidelines and are reported in the following discrete values:

$$\mathcal{Y} = \{0, 100, 200, \dots, 4900, 5000, 6000, 7000, \dots, 29000, 30000, 35000, 40000, \dots, 65000, 70000\},$$

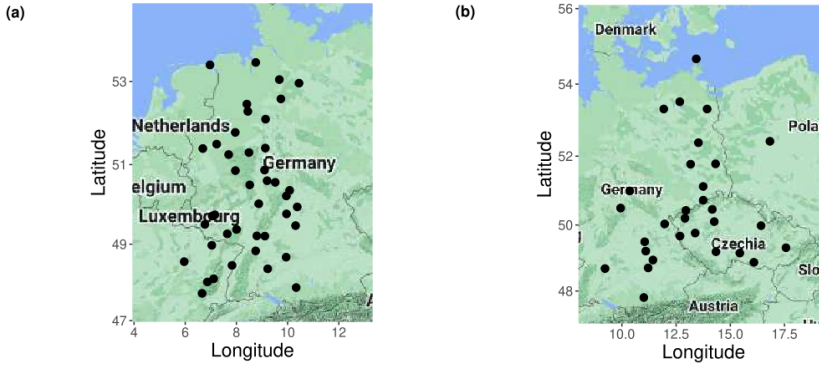


Figure 5.1: Locations of SYNOP observation stations corresponding to (a) EUPPBench benchmark dataset; (b) ECMWF forecasts for 2020-2021.

Forecasts are provided in increments of 1 m and are matched with the observations by rounding down to the nearest reported value.

First, we analyze visibility data from the *EUPPBench* benchmark dataset. This dataset includes the 51-member operational ECMWF ensemble, consisting of both the CNTL forecast and 50 interchangeable ENS members generated using random perturbations, as well as HRES forecasts. Covering the years 2017 and 2018, this dataset consists of ensemble forecasts initialized at 0000 UTC, with a forecast horizon of 120 hours and a temporal resolution of 6 hours for 42 SYNOP stations in Germany and France (see Figure 5.1a). Notably, about 1.5% of the forecast cases in this dataset are incomplete due to missing observations from certain stations. The *EUPPBench* dataset also contains gridded data, including ECMWF visibility forecasts and reforecasts. However, using reforecasts would not align with the second investigated dataset, as these are only available on specific days (Mondays and Thursdays).

In addition, we make use of a dataset comprising operational 51-member ECMWF visibility forecasts corresponding to a different region and time period. This dataset pertains to the calendar years 2020 and 2021 and includes data from 30 SYNOP stations located in Germany, the Czech Republic, and Poland (refer to Figure 5.1b). Forecasts are initialized at 0000 UTC, and we analyze 40 different lead times ranging from 6 to 240 hours, with a time step of 6 hours. This dataset is notably complete, with no missing observations, except for two specific days when certain predictions were not available: on June 2, 2021, and December 17, 2021, there were no exchangeable ensemble members for the forecast horizons of 132 hours and 114 hours, respectively.

Furthermore, we incorporate visibility forecasts from the CAMS, which serves as an independent yet comparable meteorological parameter. The de-

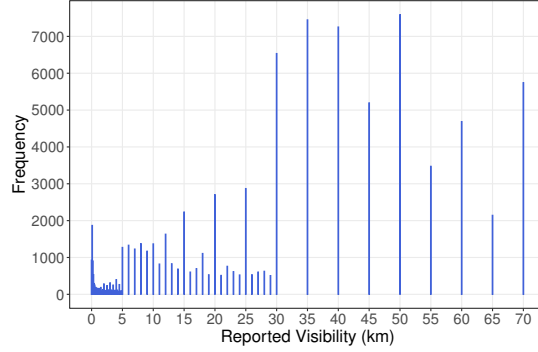


Figure 5.2: Distribution of reported visibility observations for calendar years 2020 – 2021.

terministic predictions from CAMS are part of the global atmospheric composition forecasts provided by ECMWF. The CAMS global forecast utilizes the IFS, similar to ECMWF’s weather forecasts, but incorporates specific modules to account for aerosols, reactive gases, and greenhouse gases. As the CAMS forecasts are available for the same 30 stations as the ECMWF forecasts but only up to 120 hours, with 6-hour temporal steps, in Section 5.3.2, we are limited to generating post-processed forecasts within this time frame. For further details on the calibration of the ECMWF ensemble forecasts extending to 240 hours, refer to Appendix B.

The analysis of reported visibility observations related to the ECMWF and CAMS forecasts reveals that 80.78 % of the total 87,720 observations recorded visibility of at least 10,000 m. For a visualization of the distribution of visibility observations within the dataset, please refer to Figure 5.2. The proportions of low visibility events (LVP) with observed visibility below specific thresholds are as follows: 3.19 % for dense fog (below 200 m), 5.17 % for visibility below 500 m, 6.24 % for visibility below 1000 m (fog), and 8.94 % for visibility below 3000 m. These thresholds were chosen to be consistent with recent studies focusing on the prediction of low visibility events, which are crucial in the context of aeronautical meteorology (see, for example, Penov and Guerova (2023); Shankar and Sahana (2023); He *et al.* (2024)).

5.3 Results

As mentioned in the Introduction (Section 5.1), our goal is to post-process visibility forecasts based on the forecast-observation pairs presented in Section 5.2 for our case studies. During the modelling process we consider all three

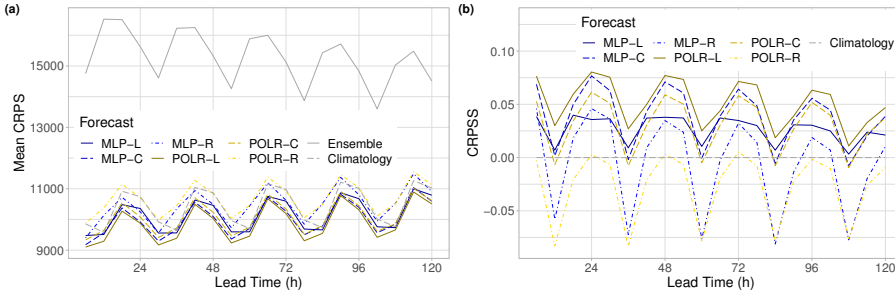


Figure 5.3: Mean CRPS of post-processed, raw and climatological EUPPBench visibility forecasts for calendar year 2018 (a) and CRPSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.

approaches to the spatial selection of training data, that is local, regional, and semi-local (see Section 3.2). In this chapter, we compare the results of Baran and Lakatos (2023) and Lakatos and Baran (2024), thereby demonstrating the usefulness of the post-processing procedures while highlighting the justification for incorporating CAMS forecasts into the post-processing workflow.

While Baran and Lakatos (2023) primarily addresses the post-processing of forecasts from the ECMWF and EUPPBench datasets, our second study explores the added value of incorporating CAMS visibility forecasts as supplementary covariates in the POLR and MLP methods for the discrete post-processing of ECMWF ensemble predictions. This approach is compared to models that rely solely on raw ECMWF ensemble data. Additionally, we evaluate the predictive performance of the resulting multivariate forecasts using ensemble copula coupling and the Schaake shuffle. In Section 5.3.1, we present the results of post-processing ensemble forecasts from the EUPPBench dataset, while Section 5.3.2 focuses on how CAMS forecasts enhance the effectiveness of the post-processing methods compared to forecasts based exclusively on the 51-member ECMWF ensemble. Detailed results on the post-processing of ECMWF ensemble forecasts can be found in Appendix B.

Since, both the MLP and POLR methods require training data, which is large enough to allow reliable modelling, and to ensure alignment with the results presented by Baran and Lakatos (2023), all MLP and POLR classifiers analyzed here employ a rolling training period of 350 days, treating each lead time independently. This, in the case of regional estimation, means 14700 forecast cases for the *EUPPBench* dataset and 10500 forecast cases for the ECMWF dataset for each training step.

We implement local (MLP-L and POLR-L), regional (MLP-R and POLR-

	MLP-L	MLP-C	MLP-R	POLR-L	POLR-C	POLR-R	Climatology
CRPS	66.35 %	65.53 %	68.47 %	64.67 %	66.05 %	69.96 %	68.19 %
LogS	45.48 %	43.91 %	44.45 %	44.86 %	43.66 %	44.55 %	64.39 %

Table 5.1: Overall mean CRPS/LogS of post-processed and climatological EUPPBench visibility forecasts for calendar year 2018 as proportion of the mean CRPS/LogS of the raw ECMWF ensemble.

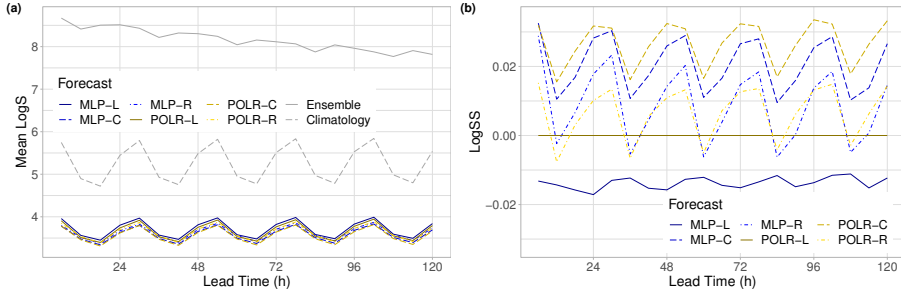


Figure 5.4: Mean LogS of post-processed, raw and climatological EUPPBench visibility forecasts for calendar year 2018 (a) and LogSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.

R), and clustering-based semi-local (MLP-C and POLR-C) training methodologies, where the latter utilizes k -means clustering of SYNOP stations. These clusters are based on three-dimensional feature vectors reflecting observed frequencies for visibility intervals of 0–5000 m, 5000–30000 m, and 30000–70000 m during the training period. With the shift of the training data, the clusters are recalculated dynamically, and in general, 8 clusters are formed in the case of the EUPPBench data and 4 in the case of forecasts for 2020–2021, provided each cluster contains at least 4 locations. Otherwise, the number of clusters is reduced and the clustering-based estimation might even fall back to regional modelling. The forecast skill of the competing post-processing methods is tested on data for a complete calendar year. As reference forecasts we consider the raw visibility ensembles and climatology, where observations of a rolling training period of a given length are considered as a forecast ensemble.

Before calculating the input features of the POLR and MLP models, a normalization of the forecasts is performed; namely, all forecast values are divided by 70000. Then we consider the following set of predictors: the (normalized) control member of the ensemble $\tilde{f}_{CNTL} := f_{CNTL}/70000$, the mean $\bar{f}_{ENS}$ of the 50 (normalized) exchangeable members, the variance s^2 of the 51-member

(normalized) operational ensemble, in the case of the EUPPBench dataset the (normalized) high-resolution forecast $\tilde{f}_{HRES} := f_{HRES}/70000$, and the proportions p_1 , p_2 and p_3 of ensemble members predicting visibility up to 1000 m, 1000 to 2000 m and more than 30000 m, respectively. The use of the control member, the mean of the exchangeable members, the ensemble variance, and possibly the high-resolution forecast as predictors is in line with the suggestions of Hemri *et al.* (2016) and rather common in parametric modelling based on ECMWF ensemble forecasts (see e.g. Gneiting, 2014), whereas covariates p_1 , p_2 and p_3 (both the number of such proportions to be considered and the boundaries) were selected based on a detailed data analysis. When analyzing the effect of CAMS predictions on the skill of classification methods, we also add the normalized CAMS forecast $\tilde{f}_{CAMS} := f_{CAMS}/70000$ to the input features and investigate the forecast skill of CAMS extended local, regional, and semi-local MLP and POLR post-processing based on the feature vector

$$(\tilde{f}_{CNTL}, \bar{f}_{ENS}, \tilde{f}_{CAMS}, s^2, p_1, p_2, p_3, \beta_1, \beta_2)^\top, \quad (5.3.1)$$

where following the suggestions of e.g. Dabering *et al.* (2017), seasonal variations are addressed by adding annual base functions

$$\beta_1(d) := \sin(2\pi d/365) \quad \text{and} \quad \beta_2(d) := \cos(2\pi d/365)$$

to the input features, where d denotes the day of the year. Note that normalization is required to have all input features commensurate, which makes modelling numerically more stable. POLR models are fit using the R package MASS (Venables and Ripley, 2002), while MLP classification is based on the package RSNNs making the Stuttgart Neural Network Simulator (Zell *et al.*, 1994) available in R. The neural network has two hidden layers with 25 – 25 neurons, the maximal number of iterations to learn is restricted to 200, the learning rate is 0.2 and both hidden layers use the logistic activation function, which parametrization is a result of extended experimentation.

Finally, extremely low predicted probabilities resulting in numerical issues in LogS calculations are handled by introducing a probability threshold $p_{\min} = 2.75 \times 10^{-5}$, which replaces all values in the predictive PMFs which are below this probability. This particular choice of $p_{\min}$ ensures that with a 1 % probability the corresponding reported visibility materializes at least once in a year; however, it is low enough, so there are no observable differences in terms of other scores between the original and the adjusted and renormalized PMFs.

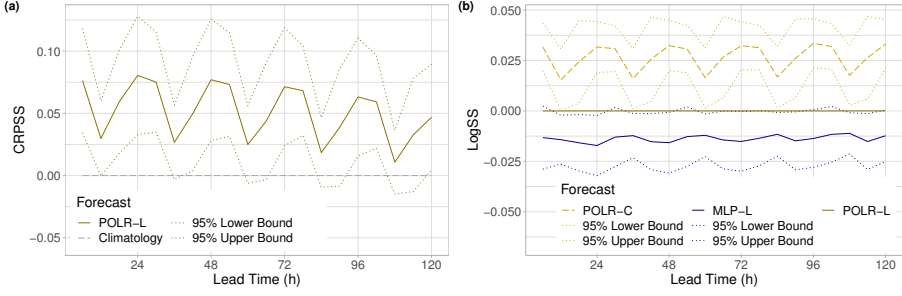


Figure 5.5: CRPS of POLR-L post-processed EUPPBench visibility forecasts for calendar year 2018 with respect to climatology (a) and LogSS of POLR-C and MLP-L approaches with respect to the POLR-L model (b) together with 95 % confidence intervals as functions of the lead time.

5.3.1 Calibration of EUPPBench visibility ensemble forecasts

The 52-member EUPPBench benchmark visibility forecasts are post-processed with the help of POLR and MLP models with a feature vector

$$(\tilde{f}_{HRES}, \tilde{f}_{CNTL}, \bar{f}_{ENS}, s^2, p_1, p_2, p_3, \beta_1, \beta_2)^\top.$$

The joint use of the high-resolution term and the operational 51-member forecasts is in line with the recommendation of the ECMWF Directorate (2012), suggesting that “the HRES and ENS forecasts should, wherever possible, be used together to provide the most detailed description of future weather and of the associated uncertainties”. Following Hemri *et al.* (2016) and Baran *et al.* (2021), for the local, semi-local and regional POLR models the coefficients of $\tilde{f}_{HRES}$, $\tilde{f}_{CNTL}$ and $\bar{f}_{ENS}$ are forced to be non-negative. Furthermore, the performance of the various forecasts is compared using forecast-observation pairs for the calendar year 2018. For the reference climatology a 40-day rolling training period is chosen. This training period length is short enough to allow climatological forecasts to adapt to seasonal variations in visibility. Finally, we tested the skill of those climatological forecasts, that are based on the same 350-day training period as the post-processing approaches.

Both climatological and post-processed forecasts (MLP and POLR models) result in far lower mean CRPS values than the raw EUPPBench ensemble for all lead times (see Figure 5.3a). As depicted in Figure 5.3b, climatology performs surprisingly well, clearly outperformed for all forecast horizons only by the locally trained POLR and MLP approaches. Except for lead times corresponding to 1200 UTC, POLR-C and MLP-C also yield positive CRPS with respect to

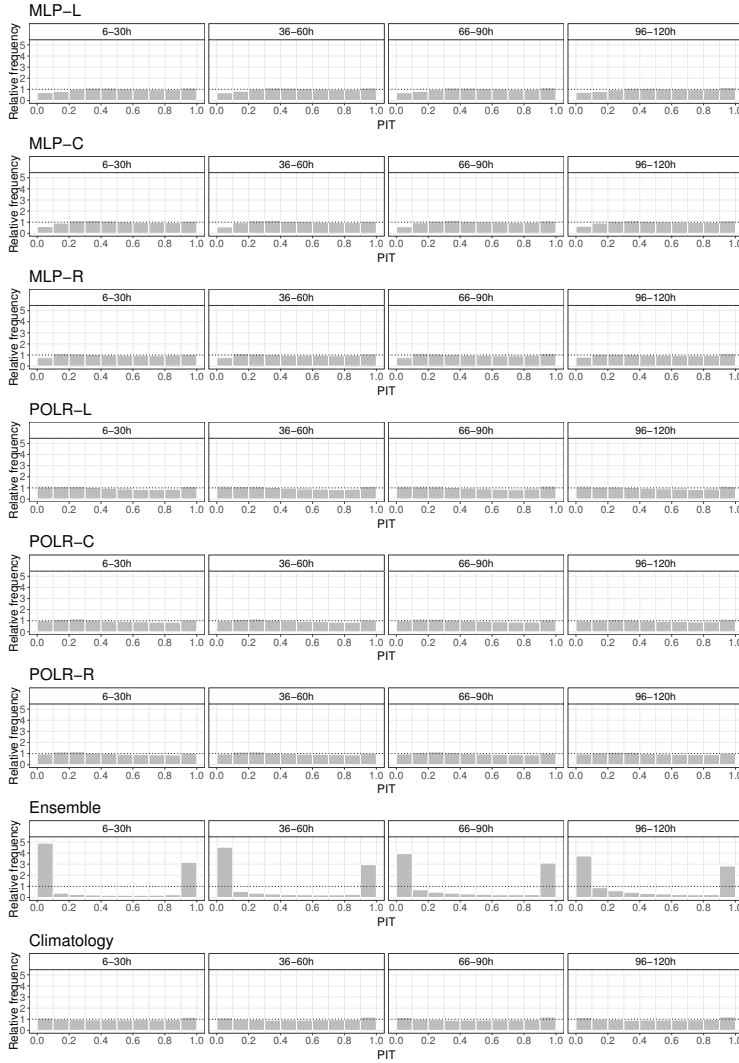


Figure 5.6: PIT histograms of post-processed, climatological and raw visibility EUPPBench forecasts for calendar year 2018 for lead times 6–30 h, 36–60 h, 66–90 h and 96–120 h.

climatology and are superior to MLP-L, whereas POLR-R and MLP-R follow a similar pattern but perform much worse than the corresponding semi-local models. The ranking of forecasts based on Figure 5.3b (POLR-L – MLP-C –

MLP-L	MLP-C	MLP-R	POLR-L	POLR-C	POLR-R
2.19×10^{-13}	1.80×10^{-14}	2.82×10^{-4}	5.55×10^{-18}	1.15×10^{-6}	4.57×10^{-4}

Table 5.2: Overall mean p -values of the α_{1234}^0 tests for uniformity of the PIT values for calendar year 2018.

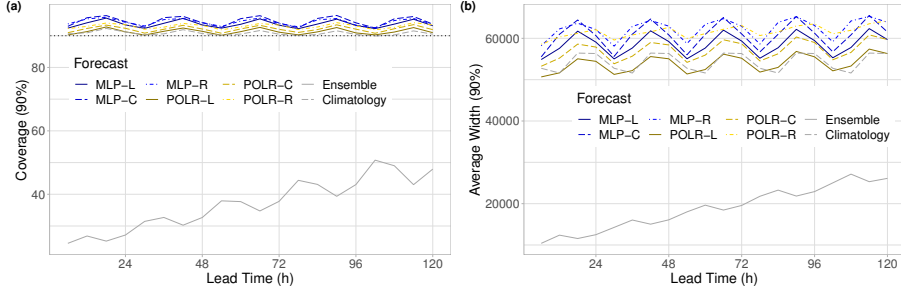


Figure 5.7: Coverage (a) and average widths (b) of 90% central prediction intervals of raw and post-processed EUPPBench visibility forecasts for calendar year 2018 as functions of the lead time. In panel (a) the ideal coverage is indicated by the horizontal dotted line.

POLR-C – MLP-L – Climatology – MLP-R – POLR-R) is also confirmed by Table 5.1 reporting the overall improvement in mean scores with respect to the raw EUPPBench visibility ensemble forecasts.

In terms of the mean LogS, climatology is far behind post-processed forecasts. According to Figure 5.4b, semi-local models clearly outperform their local counterparts and except for lead times corresponding to 1200 UTC, even the regional approaches are superior to the local ones. From the studied calibration methods POLR-C results in the lowest mean LogS, closely followed by the MLP-C, whereas MLP-L shows the poorest performance, see also Table 5.1.

The good performance of climatology with regard to the mean CRPS is also confirmed by Figure 5.5a, as at forecast horizons 36 h, 60 h, 66 h, 84 h, 90 h, 108 h and 114 h even the best performing POLR-L model fails to result in significantly positive CRPSS at a 5% level. Furthermore, the LogSS of POLR-C with respect to POLR-L is significantly positive for all investigated lead times, whereas the 95% upper bound for the LogSS of the least performing MLP-L almost coincides with the reference line. This means a barely significant difference between the MLP-L approach and POLR-L.

Based on the PIT histograms of Figure 5.6, raw EUPPBench forecasts are

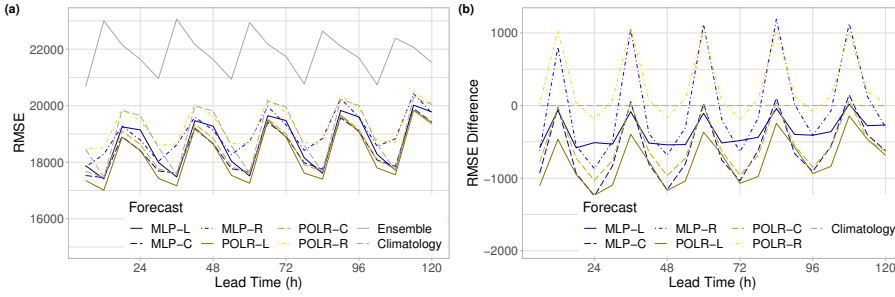


Figure 5.8: RMSE of the mean EUPPBench visibility forecasts for calendar year 2018 (a) and difference in RMSE from climatology (b) as functions of the lead time.

highly underdispersive with a small bias increasing with the forecast horizon, whereas climatology and all post-processing approaches display rather flat histograms indicating improved calibration. Nevertheless, MLP predictions exhibit a tiny overdispersion and climatological forecasts are slightly underdispersive. For all calibrated forecasts, the moment-based α_{1234}^0 test (Knüppel, 2015) rejects uniformity at a 5% level of significance for all 20 investigated lead times. Nevertheless, a possible ranking of the competing post-processing approaches in terms of goodness of fit of PIT can be obtained from Table 5.2 reporting the mean p-values of the α_{1234}^0 tests over all considered lead times (the larger the better). Note that for raw and climatological forecasts, the PIT values are concentrated around 53 and 41 bins (ensemble size +1), respectively. Hence, these PIT histograms fall back to the corresponding verification rank histograms (Wilks, 2019, Section 9.7.1).

According to Figure 5.7a, with the increase of the forecast horizon the coverage of 90 % central prediction intervals of the raw EUPPBench forecasts increases from 22 % to 51 %, and shows a clear diurnal cycle. This diurnal cycle is also preserved in the coverage values of climatological and post-processed forecasts; however, these curves do not display any further dependence on the forecast horizon and are just slightly above 90 %. The best coverage again corresponds to climatological forecasts with a mean absolute deviation from the nominal value of 0.82 %, followed by the POLR models (POLR-L: 1.32 %; POLR-C: 2.12 %; POLR-R: 3.03 %) and MLP approaches (MLP-L: 3.69 %; MLP-C: 4.59 %; MLP-R: 4.35 %). The average widths of 90 % central prediction intervals depicted in Figure 5.7b are nicely in line with the corresponding coverage values, with the POLR-L approach providing the sharpest post-processed forecasts.

Furthermore, all post-processing methods result in a considerable decrease

in the RMSE of the mean forecast (see Figure 5.8a). Moreover, according to Figure 5.8b, for the EUPPBench data climatology is fully able to catch up with the POLR and MLP models and for forecast horizons corresponding to 1200 UTC it definitely outperforms their local versions.

Finally, 350-day climatology performs well only in terms of the LogS; however, even with respect to this particular score it is still slightly behind the post-processed forecasts for all forecast horizons. The overall mean CRPS of 350-day climatological forecasts is 73.67 % of the mean CRPS of the raw EUPPBench ensemble (compare with the corresponding values of Table 5.1), its mean absolute deviation in 90 % coverage from the nominal value is 3.19 %, and the sharpness of the corresponding central prediction interval is between the sharpnesses of locally and semi-locally trained MLP models.

5.3.2 Post-processing ECWMF visibility ensemble forecasts using CAMS covariates

In this subsection, we focus on the post-processing of ECMWF visibility ensemble forecasts, integrating CAMS visibility forecasts as additional covariates. Our goal is to determine whether the inclusion of CAMS data enhances forecast accuracy compared to models based exclusively on ECMWF ensemble predictions. By incorporating CAMS forecasts—specifically designed to account for atmospheric elements like aerosols and reactive gases—we also aim to achieve a more precise representation of low-visibility events.

Univariate performance

To simplify the presentation, the predictive performance of univariate visibility forecasts is evaluated with the help of the mean CRPS and mean LogS over all forecast cases in the verification period, together with the corresponding skill scores. Furthermore, their skill in predicting low-visibility events is quantified with the help of the mean BS. In the case of post-processed forecasts based solely on ECMWF ensemble predictions, further evaluation metrics can be found in Appendix B. Reference MLP and POLR models are now based on the feature vector

$$(\tilde{f}_{CNTL}, \bar{f}_{ENS}, s^2, p_1, p_2, p_3, \beta_1, \beta_2)^\top, \quad (5.3.2)$$

where for the latter the weights of $\tilde{f}_{CNTL}$ and $\bar{f}_{ENS}$ are again forced to be non-negative. The same constraint is applied for the coefficient $\tilde{f}_{CAMS}$ of the CAMS extended local, semi-local and regional POLR models with input feature vector (5.3.1).

As demonstrated in Section 4.2 of Baran and Lakatos (2023), in terms of the mean CRPS and mean LogS (see Figures B.1a and B.2a in Appendix B), up to 120 h, all MLP and POLR post-processed forecasts have shown significant

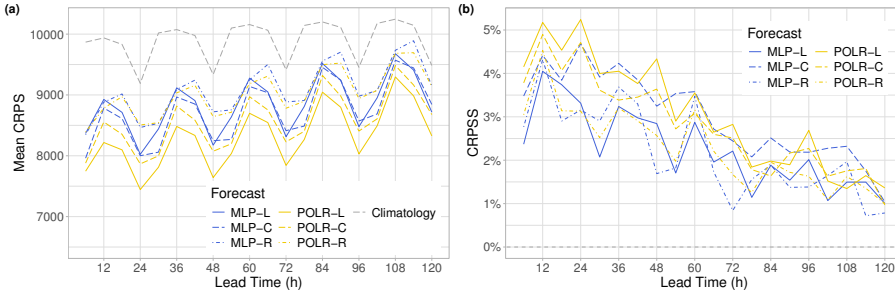


Figure 5.9: Mean CRPS of climatological and post-processed visibility forecasts based on CAMS extended MLP and POLR models (a) and CRPSS of CAMS extended post-processed forecasts with respect to the corresponding MLP and POLR models based only on ECMWF predictions (b) as functions of the lead time.

improvement compared to both the raw ensemble and climatology, which is now based on the observations of the past 30 days. The extent of this improvement is particularly striking in the case of the raw ensemble, which we have chosen to exclude from our univariate analysis. Figure 5.9a displays the mean CRPS of climatological and CAMS extended MLP and POLR forecasts as functions of the forecast horizon. All post-processed predictions outperform climatology for all lead times, and the ranking of the various approaches is identical to the one for the corresponding reference models based on feature vector (5.3.2), see Figure 2 of Baran and Lakatos (2023) or Figure B.1a in Appendix B. The advantage achieved in mean CRPS by utilizing the CAMS extended feature vector (5.3.1) over models based solely on ECMWF forecasts is illustrated by the CRPSS values of Figure 5.9b. In general, it can be observed that the integration of CAMS into the post-processing workflow for regional models yields the lowest benefits, but this advantage is still noteworthy. Note that the best performing POLR and MLP variants, which are the local POLR model and the semi-local MLP, profit the most from the extension of the input features with an average CRPSS of around 3%. One should also remark that the improvement in mean CRPS from including CAMS predictions might further decrease for forecast horizons beyond 120 h.

The mean LogS values of Figure 5.10a tell the same story about the various post-processing approaches as Figure 3 of Baran and Lakatos (2023), reproduced as Figure B.2 in Appendix B. Adding the CAMS forecast to the input features of the local, semi-local, and regional MLP and POLR models does not change the ranking of the forecasts; however, as portrayed in Figure 5.10b, it results in a modest 0.4 – 0.9 % average improvement in LogS.

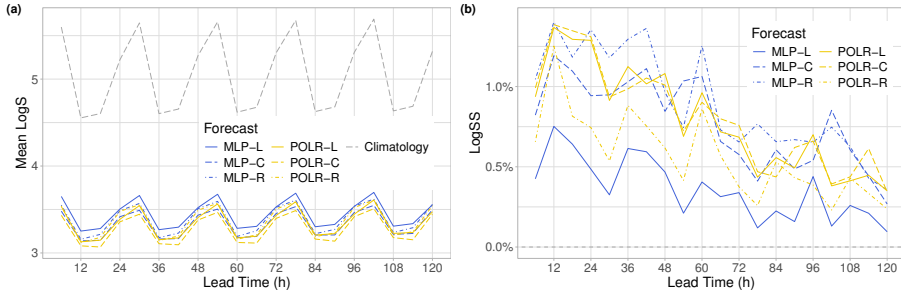


Figure 5.10: Mean LogS of climatological and post-processed visibility forecasts based on CAMS extended MLP and POLR models (a) and LogSS of CAMS extended post-processed forecasts with respect to the corresponding MLP and POLR models based only on ECMWF predictions (b) as functions of the lead time.

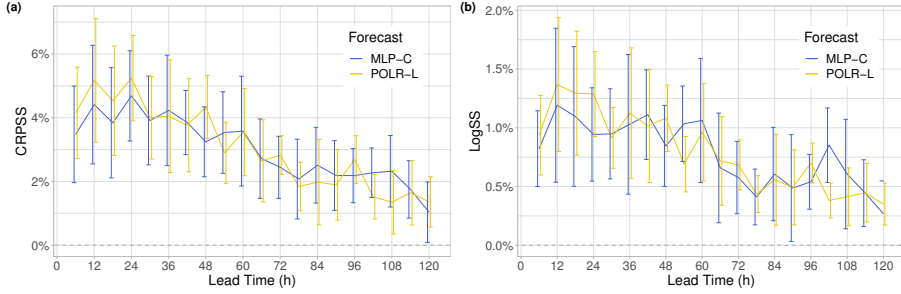


Figure 5.11: CRPSS (a) and LogSS (b) of CAMS extended MLP-C and POLR-L forecasts with respect to the corresponding models based only on ECMWF predictions together with 95 % confidence bounds as functions of the lead time.

To address the significance of gain of CAMS extended post-processing models with respect to the corresponding forecasts based only on ECMWF predictions, in Figure 5.11 we accompany the CRPSS and LogSS of the MLP-C and POLR-L approaches, which are the best model variants according to Baran and Lakatos (2023), by 95 % block-bootstrap confidence intervals (for details see Figure B.3 in Appendix B). As depicted in Figures 5.11a and 5.11b, up to the maximal studied lead time of 120 h, the improvement in terms of both scores is significant at a 5 % level.

Finally, we investigate the skill of the best-performing CAMS extended MLP-C and POLR-L methods for the visibility range considered in aeronautical

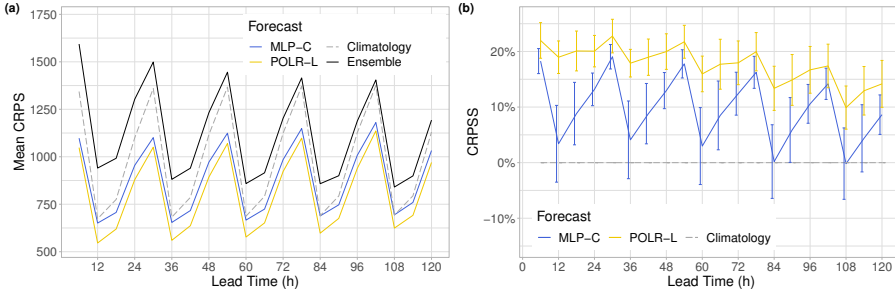


Figure 5.12: Mean CRPS of raw, climatological and CAMS extended MLP-C and POLR-L forecasts censored from above at 10 km (a) and CRPSS of CAMS extended post-processed forecasts with respect to climatology (b) together with 95 % confidence bounds as functions of the lead time.

meteorology. We censor the observed classes from above at 10 km, so reported observations exceeding this value are rounded down to 10 km, resulting in 56 different values, while the corresponding probabilities are pooled. Naturally, in this case, the definition of the CRPS should be modified accordingly.

As depicted in Figure 5.12a, censoring at 10 km does not change the ranking of the different forecasts in terms of the mean CRPS. However, compared to Figure 2 of Baran and Lakatos (2023), –reproduced as Figure B.1 in Appendix B–, the advantage of climatology over the (CAMS extended) raw ensemble is much smaller, and the difference between climatological and post-processed forecasts is also reduced compared to Figure 5.9a. Nevertheless, the local POLR model significantly outperforms climatology for all lead times with a CRPSS averaging 17.7 % (Figure 5.12b), whereas for the semi-local MLP approach, the average gain in CRPSS over climatology is 9.5 %, and the difference fails to be significant at a 5 % level for forecasts corresponding to 1200 UTC.

The Brier skill scores of Figure 5.13 assessing the relative performance of post-processed forecasts with respect to climatology for various low visibility events convey the same message. The advantage of the POLR-L approach is significant in all four cases for all lead times with BSS values between 9.3 % and 21.4 %, while the MLP-C method again cannot show a significantly positive skill for 1200 UTC observations.

Multivariate post-processing

Continuing our analysis, we delve into the comparison of forecast skill among multivariate approaches detailed in Chapter 2. In the case of the applied multivariate methods, calibrated samples of size 51 are employed, mirroring the di-

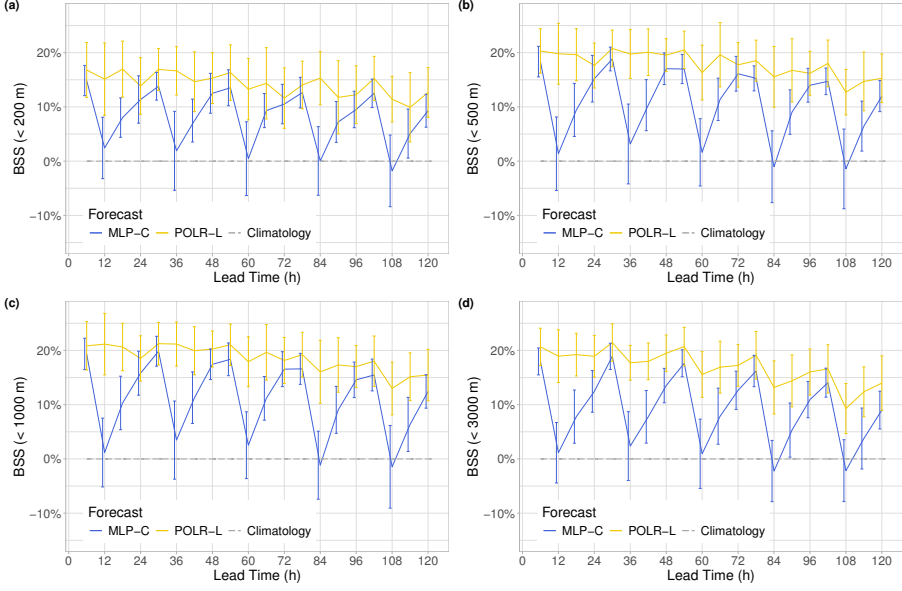


Figure 5.13: BSS of CAMS extended MLP-C and POLR-L forecasts with respect to climatology for the events where visibility is below 200 m (a), 500 m (b), 1000 m (c), and 3000 m (d) together with 95% confidence bounds as functions of the lead time.

mentality of the raw ECMWF ensemble. When employing the SSh method, which relies on historical data to establish dependency templates, we utilize observations obtained from the rolling training window applied during univariate calibration. This decision is made based on the finding of Lakatos *et al.* (2023), that extending these pool dates to encompass the entire available historical dataset may only yield negligible advantages. In our case, this entails selecting 51 days from a pool of 350 dates (which is the length of the training period). As mentioned, in both the ECC and SSh methods, the initial step involves generating samples from the calibrated PMFs, which we perform by considering their equidistant quantiles.

To simplify presentation, as univariate calibrated forecasts we consider the CAMS extended MLP-C and POLR-L predictions and their counterparts based only on the ECMWF ensemble. The corresponding multivariate predictions obtained with the help of the ECC and SSh are referred to as MLP-C ECC, POLR-L ECC, and MLP-C SSh, POLR-L SSh, respectively. We also report the performance of naive multivariate forecasts obtained by simply arranging the

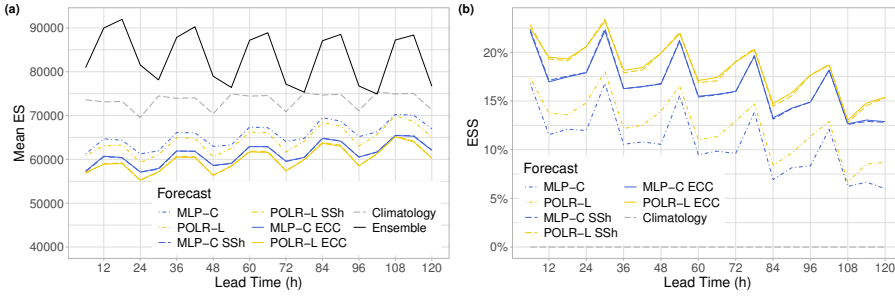


Figure 5.14: Mean ES of raw, climatological and post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models (a) and ESS of CAMS extended post-processed forecasts with respect to multivariate climatology (b) as functions of the lead time.

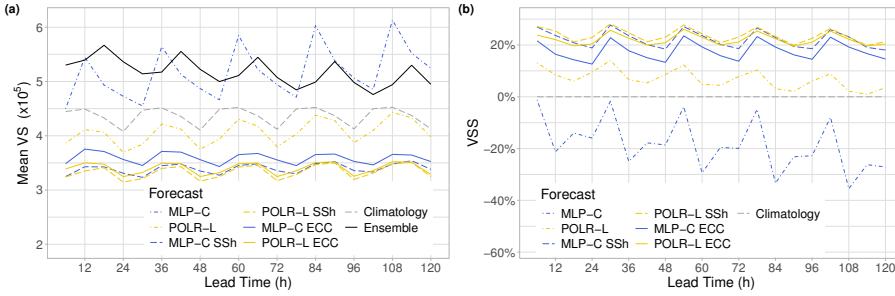


Figure 5.15: Mean VS of raw, climatological and post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models (a) and VSS of CAMS extended post-processed forecasts with respect to climatological climatology (b) as functions of the lead time.

51-member station-specific MLP-C and POLR-L samples into 30-dimensional vectors. In the multivariate context, notations MLP-C and POLR-L will refer to these "independent" predictions. Finally, as a multivariate climatological forecast, we consider the vector of 51 equidistant quantiles of 30-day climatological PMFs corresponding to the investigated SYNOP stations. This approach results in 51 continuous 30-dimensional vectors. In principle, one could also consider the vectors of station-specific 51-day climatological forecasts; however, due to their discrete nature (all vector coordinates are elements of the discrete set $\mathcal{Y}$), they would substantially differ from the competing calibrated

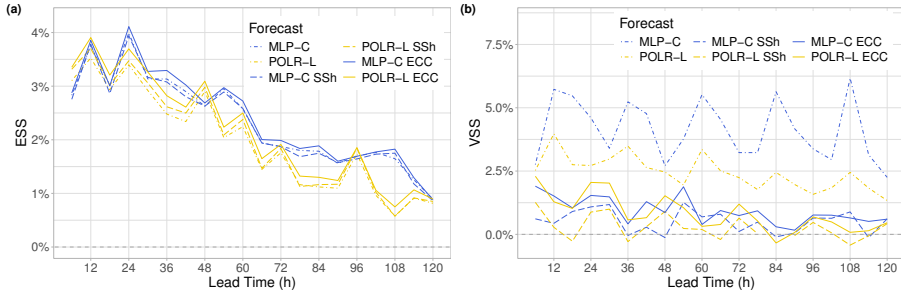


Figure 5.16: ESS (a) and VSS (b) post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models with respect to the corresponding forecasts based only on ECMWF predictions as functions of the lead time.

predictions.

First, we assess the performance of the previously discussed multivariate forecasts based on the energy score. Figure 5.14a clearly illustrates the substantial superiority of each CAMS extended multivariate forecast over the raw ensemble and the 30-day climatology in terms of this scoring rule. Notably, the semi-local MLP ECC and local POLR ECC exhibit the best predictive performance, extremely closely approaching their SSh counterparts, thereby minimizing performance disparities and highlighting the success of multivariate post-processing. According to the energy skill scores of Figure 5.14b, the univariate MLP-C and POLR-L models alone demonstrate a substantial average improvement of 12 % over multivariate climatology, while the multivariate models show even more pronounced performance enhancements, averaging 17 %. This analysis emphasizes that, at least in our dataset, irrespective of the selected dependency template, multivariate models consistently exhibit improvement compared to independent calibration and climatology.

Figure 5.15a presents a slightly different picture in terms of model ranking, which might be a natural phenomenon when considering the variogram score depicted here. This verification measure clearly illustrates the success of multivariate post-processing in the redefinition of covariance structures. As expected, (CAMS extended) naive models, such as MLP-C and POLR-L, do not capture the spatial dependencies and might underperform the raw forecast vectors and multivariate climatology, as well. According to the VSS values of Figure 5.15b, the two-step multivariate models demonstrate a notable average improvement of over 17.6 % with respect to the reference climatological predictions. Among these approaches, the POLR variants exhibit the highest skill, particularly the POLR-L SSh configuration. Moreover, it is noteworthy

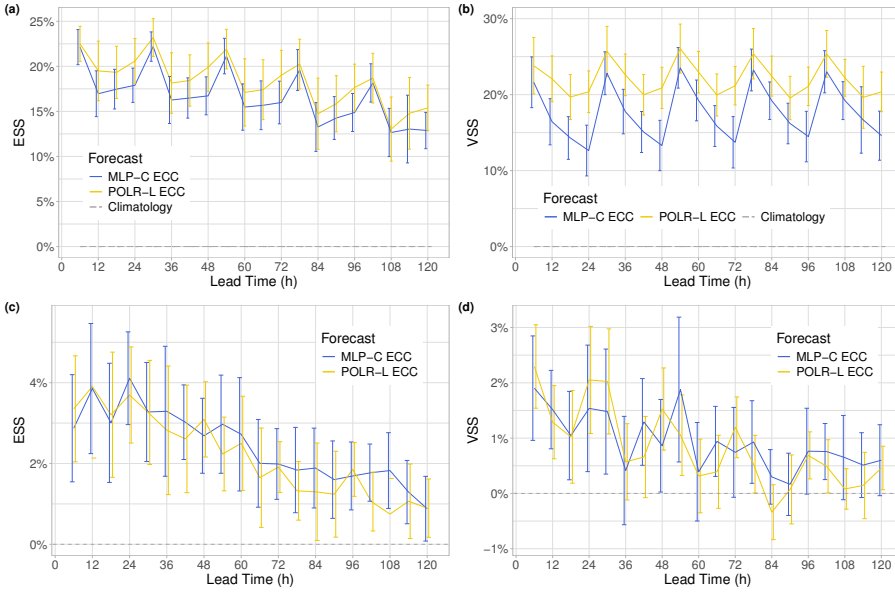


Figure 5.17: ESS (a,c) and VSS (b,d) of CAMS extended MLP-C ECC and POLR-L ECC multivariate forecasts with respect to multivariate climatology (a,b) and to the corresponding MLP-C ECC and POLR-L ECC forecasts based only on ECMWF predictions (c,d) together with 95 % confidence bounds as functions of the lead time.

that a more pronounced advantage was attained with the SSh models obtaining dependence templates from past observations, compared to the ECC forecasts utilizing the rank structure of raw ensemble.

In Figure 5.16, we illustrate how the CAMS extended post-processed forecasts compare to the corresponding multivariate predictions from classifiers that ignore CAMS forecasts. Both the ESS and VSS reveal a diminishing advantage of CAMS integration with increasing forecast horizons. In the case of the ES the differences between the various post-processed forecasts are minor (see Figure 5.16a) with the MLP-C ECC variant showing the most overall improvement. The VSS values of Figure 5.16b again tell a different story. With independent MLP-C and POLR-L models, there is an average improvement of around 4 % and 2.5 %, respectively, in variogram scores suggesting that using CAMS as a feature helps in better preserving spatial dependencies. For the multivariate models, the gain is much smaller, and there are several lead times where some models display even negative skill.

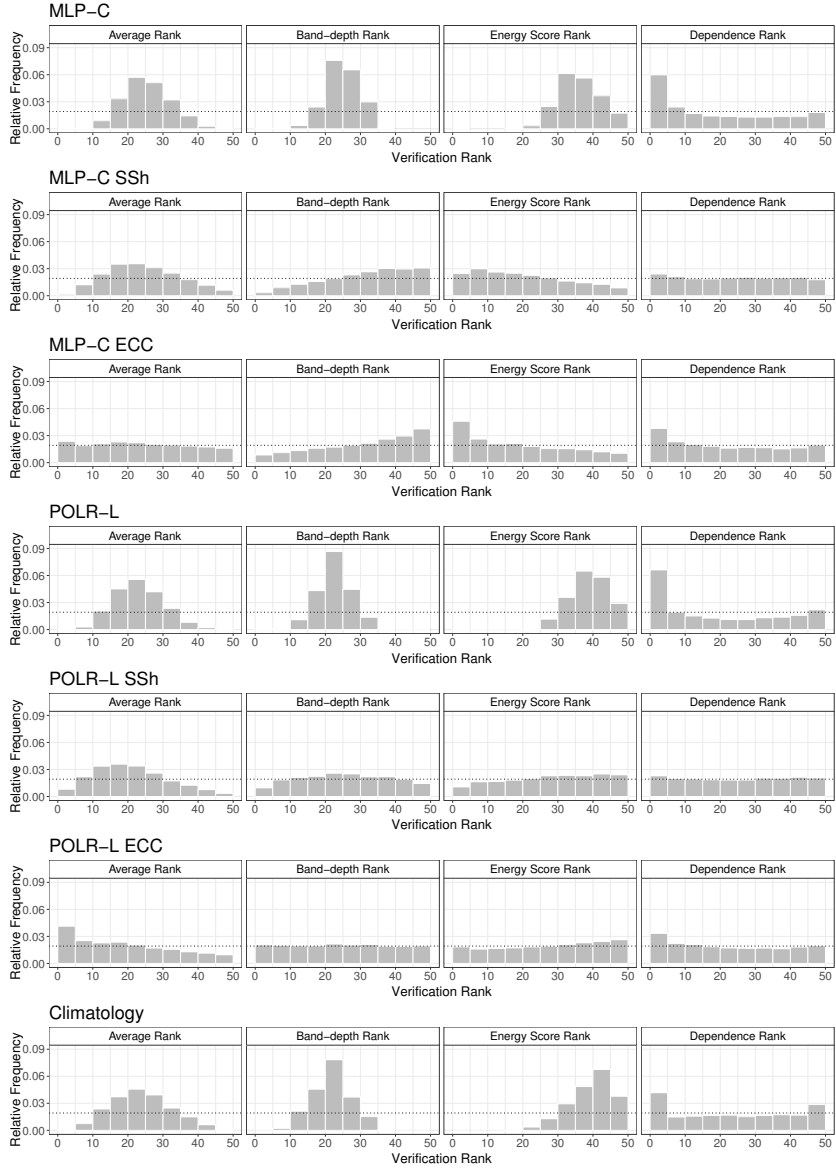


Figure 5.18: Rank histograms of climatological and post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models.

For better clarity, in Figure 5.17, we accompanied the ESS and VSS values of MLP-C ECC and POLR-L ECC forecasts – which are overall considered the best performers – with 95 % bootstrap confidence bounds. In cases where climatology serves as the reference, as seen in Figures 5.17a and 5.17b, in general, the local POLR-based ECC outperforms its MLP-C counterpart. However, for forecasts corresponding to 0600 UTC, both in terms of ES and VS, the difference is not significant, indicating that the advantage of the POLR-L ECC approach over the MLP-C ECC is more apparent for the other three studied observation times (1200, 1800, 2400 UTC). Furthermore, Figures 5.17c and 5.17d demonstrate that incorporating CAMS forecasts as an additional feature leads to a significant enhancement in ES for all investigated lead times, whereas for VS there are a few instances of longer lead times exhibiting negative skill within the confidence bounds.

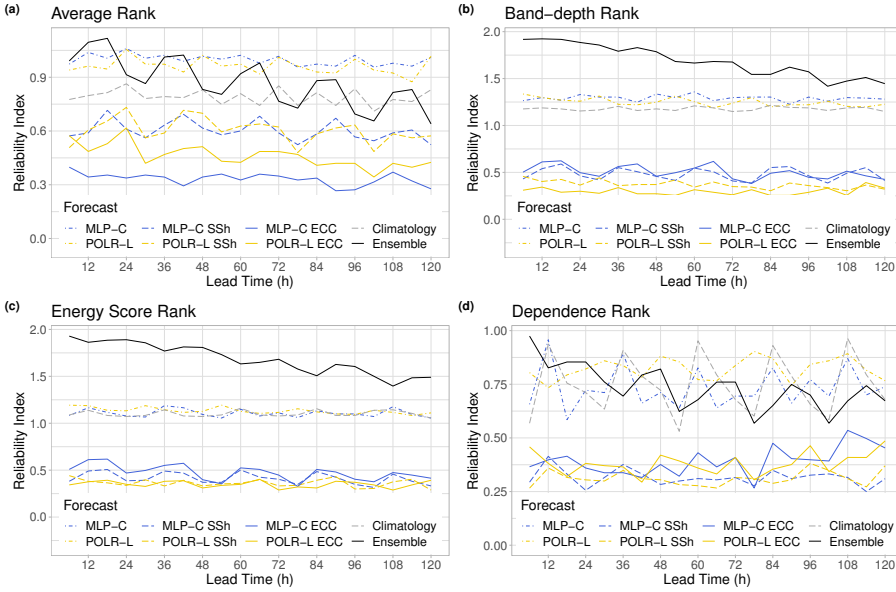


Figure 5.19: Reliability indices of rank histograms of raw, climatological and post-processed multivariate visibility forecasts based on CAMS extended MLP-C and POLR-L models as functions of the lead time.

In addition to the previous analyses, we also generated multivariate rank histograms to thoroughly assess the differences in calibration of the individual forecasts (see Figure 5.18). In general, properly calibrated forecasts result in flat histograms, and the deviation from uniformity can be quantified by the

reliability index (3.1.3). In principle, one can have an instant ranking of the various multivariate predictions with the help of the reliability indices, displayed in Figure 5.19 (the smaller the better); however, the interpretation of the different rank histograms can vary depending on the functions applied to determine the pre-ranks, as these functions may be more sensitive to detecting different deficiencies. Therefore, as argued by Thorarinsdottir *et al.* (2016) and Allen *et al.* (2024), using multiple pre-rank functions to evaluate the performance of the forecasts could be beneficial, providing a more comprehensive comparison. In the case of naive MLP-C and POLR-L models, hump-shaped average- and band-depth rank histograms indicate overestimated correlations, which is also supported by the skewed energy score- and dependence rank histograms. To a smaller extent, the same applies to the multivariate climatology. Multivariate calibration results in a substantial improvement, especially in the case of MLP-C ECC and POLR-L ECC forecasts. According to Figure 5.19, in terms of the average ranks, these two methods provide by far the lowest reliability indices and are very competitive with the MLP-C SSh and POLR-L SSh approaches in the three other cases, as well.

5.4 Conclusions

This chapter synthesizes findings from two papers that focus on the statistical post-processing of visibility forecasts using distinct datasets. The first study explores univariate discrete post-processing of visibility forecasts from the ECMWF ensemble and the EUPPBench dataset, evaluating the performance of proportional odds logistic regression (POLR) and multilayer perceptron (MLP) classifiers. We analyze 51-member operational ECMWF forecasts across 30 locations in Central Europe, with a forecast horizon of 240 hours and a time step of 6 hours for the years 2020 and 2021. Results indicate that both climatological and post-processed forecasts substantially outperform raw ensemble forecasts in terms of mean CRPS, mean LogS, and RMSE, though the performance gap narrows as the forecast horizon extends (see Appendix B for ECMWF results, and Section 5.3.1 for EUPPBench outcomes). POLR models demonstrate superior performance compared to MLP counterparts, echoing the findings of Baran *et al.* (2021) regarding total cloud cover forecasts. Notably, the locally trained POLR model excels among post-processing methods, while semi-local training often outperforms local training in machine learning approaches.

The second study continues from Baran and Lakatos (2023), investigating location-specific discrete post-processing of visibility. We address three main questions: the capability of advanced multivariate post-processing approaches to restore spatial dependencies lost during separate calibration of visibility ensemble forecasts; the impact of incorporating CAMS visibility predictions

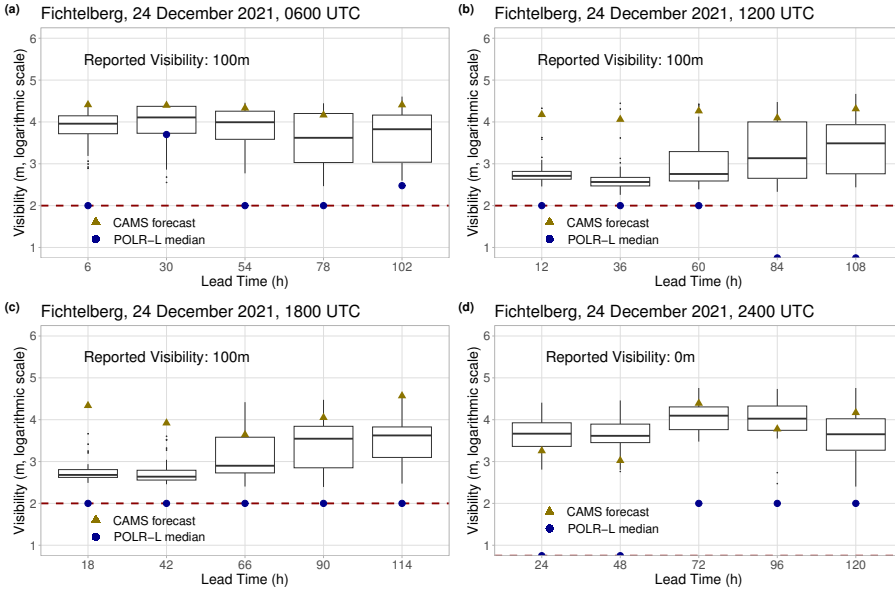


Figure 5.20: Boxplots of ECMWF visibility forecasts (logarithmic scale) for Fichtelberg mountain (Germany) for 0060 UTC (a), 1200 UTC (b), 1800 UTC (c), and 2400 UTC on the Christmas Eve of 2021 together with the corresponding CAMS forecasts and the median forecasts of the CAMS extended POLR-L model. The observed visibility is indicated by the horizontal dashed line. Values on the bottom line of panels (b) and (d) correspond to 0 m visibility.

as additional covariates on forecast skill; and the predictive performance of post-processed forecasts in aeronautical meteorology.

For the top-performing univariate approaches (MLP-C and POLR-L), integrating CAMS predictions leads to significant improvements of 1% to 5.25% in mean CRPS and 0.27% to 1.37% in mean LogS, with diminishing gains as lead time increases.

To exemplify the effect of statistical post-processing, we consider Christmas Eve 2021 at Fichtelberg mountain, Germany, where reported visibility reached a maximum of 100 m. Figure 5.20 shows that ECMWF forecasts were accurate only at specific times and short lead times; CAMS predictions underperformed, while the median CAMS extended POLR-L forecast proved highly accurate.

This approach effectively predicts low visibility events crucial for aeronautical applications, with average improvements in BS for conditions below 200 m, 500 m, 1000 m, and 3000 m ranging from 14.22% to 18.59%.

We performed joint multivariate post-processing for all 30 SYNOP observation stations using two-step ECC and SSh methods, leveraging the dependence structure of the raw vector ensemble forecasts and historical observations. Both methods significantly outperform reference independently calibrated multivariate forecasts. When compared to forecasts based solely on ECMWF predictions, incorporating CAMS yields an additional improvement of 0.58% to 4.11% in energy score across all models. In terms of correlation specification, the CAMS extended forecasts surpass climatology by over 20% for ECC and SSh methods, despite a 5% advantage for independent reference methods. Among the competing multivariate predictions, the ECC-corrected MLP-C and POLR-L methods exhibit the best overall performance.

Our findings highlight the value of CAMS visibility predictions as additional covariates in discrete post-processing of visibility ensemble forecasts, demonstrating the effectiveness of simple two-step multivariate methods in capturing spatial dependencies for this weather variable.

While this study focuses on a specific geographical area and time frame, we believe the conclusions regarding the skill of post-processed forecasts are broadly applicable. Results from Baran and Lakatos (2023) affirm that the proposed MLP and POLR approaches also perform well across different ensemble domains and time periods, with the two-step multivariate methods proven effective in various weather scenarios. Notably, a variant of ECC developed for hourly rainfall prediction is currently operational at Météo France (Taillardat and Mestre, 2020).

The classification-based univariate post-processing methods are limited to discrete observations, necessitating more training data with increased visibility categories. However, with the state-of-the-art visibility sensors offering 1 m resolution—matching that of ECMWF and CAMS forecasts—visibility can be treated as a continuous quantity. Future research should explore the use of CAMS forecasts and other visibility-related covariates in continuous post-processing models. Potential parametric candidates include the BMA approach of Chmielecki and Raftery (2011), which utilizes a mixture of beta distributions for predictive visibility, and the more recent censored gamma and censored truncated normal mixture model proposed by Baran and Baran (2024). A natural extension of this work would involve adapting the classification and interpolation-based approach of Scheuerer *et al.* (2020) to visibility forecasts.

Chapter 6

Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts

As demonstrated in the previous chapters, univariate post-processing methods, such as EMOS, BMA, and the classification techniques used for visibility predictions, can effectively calibrate forecasts for various quantities. However, these approaches have a notable drawback: they assume statistical independence among different forecast dimensions, such as lead times, locations, and weather variables, which can hinder the generation of realistic forecast scenarios.

End-users often require coherent predictions that account for dependencies, such as total precipitation across a hydrological catchment or the interaction between precipitation and temperature, particularly near freezing conditions. Consequently, there is a need for explicitly re-establishing dependence structures through multivariate post-processing techniques. This chapter aims to evaluate and compare several state-of-the-art methods for multivariate post-processing, focusing on real NWP ensemble forecasts and observations from the ECMWF.

To address this need, various copula-based approaches have been developed. Most advanced multivariate post-processing applications utilize an empirical copula derived from a dependence template based on either NWP ensembles or historical observations. The former is known as ensemble copula coupling (ECC), while the latter is referred to as the Schaake shuffle (SSh), both of which

were discussed in Chapter 2 and utilized in the previous chapter. For a comprehensive discussion on copula-based methods for incorporating dependence structures, see Schefzik and Möller (2018).

Lerch *et al.* (2020) conducted an extensive simulation study comparing different variants of ECC, SSh, and the parametric Gaussian copula approach (GCA). Their findings indicated that the benchmark methods, specifically ECC-Q (ECC with equidistant quantile sampling) and SSh, generally performed well. Additionally, they suggested that more sophisticated methods, such as dual ECC (dECC) or GCA, offer substantial predictive benefits only under specific conditions. In this study, we investigate whether these findings hold true when using real NWP ensemble forecasts and observations. In order to ensure that our results provide an analysis as comparable as possible to that of Lerch *et al.* (2020), we refrained from using alternative approaches in our case studies, such as one-step models like the member-by-member post-processing techniques (Van Schaeybroeck and Vannitsem, 2015) or the direct modelling of multivariate distributions (Baran and Möller, 2015).

To this end, we apply univariate EMOS in conjunction with a range of ECC and SSh variants to NWP ensemble forecasts of temperature, wind speed, and precipitation provided by the ECMWF, verifying against observations from SYNOP stations. To the best of our knowledge, this work is the first to compare a broad array of state-of-the-art two-step methods for multivariate post-processing, including recently proposed similarity-based Schaake shuffle approaches and dual ECC.

While we concentrate on the multivariate post-processing step, we implement univariate calibration following the EMOS methods outlined by Hemri *et al.* (2014). For the multivariate step, we compare three naive approaches derived directly from univariate EMOS with four variants of ECC and five variants of SSh, focusing specifically on temporal dependence for clarity and simplicity.

6.1 Data

We assess a range of multivariate post-processing techniques applied to global ECMWF ensemble forecasts of 2m temperature (T2M), 10m wind speed (V10), and 24-hour precipitation accumulation (PPT24) from January 1, 2002 to March 20, 2014. The datasets used are the same as those analyzed in Hemri *et al.* (2014) and include the ECMWF HRES forecasts, the 50-member operational ensemble generated through random perturbations, and the CNTL run initialized at 1200 UTC, with forecast horizons ranging from 1 to 10 days, alongside the corresponding observations. After an initial quality control process to eliminate SYNOP stations with missing or irregular data, we retained 4,160, 4,388, and 2,917 stations worldwide for T2M, V10, and PPT24, respectively.

For further details regarding the datasets and the quality control procedures employed, refer to Hemri *et al.* (2014).

6.2 Results

The predictive performance of the different multivariate post-processing methods is investigated in three case studies based on global ECMWF T2M, V10 and PPT24 ensemble forecasts. To assess calibration of probabilistic forecasts we use ES, the VS of order 1 (for which constant weights with $\omega_{ij} = 1$ are used) and we also investigate the multivariate rank histograms together with the corresponding reliability indices, whereas multivariate point forecasts are evaluated with the help of the mean Euclidean error (EE). Note that in addition we studied the regularized Dawid-Sebastiani scores (RDSS; Wilks, 2020) of the raw ensemble and multivariate forecasts addressing temporal dependence. However, for our data sets the RDSS results in a similar ranking of the various post-processing methods as the ES. In the interest of brevity, we restrict our attention to the ES and VS here, and provide more detailed results for the RDSS in Appendix C.

While our focus is on the multivariate predictive performance, the first step in each of the multivariate methods described in Chapter 2 is the independent calibration of ensemble forecasts with different forecast horizons. Hence, we start with a brief description of the details of univariate post-processing, followed by the results for the multivariate models.

6.2.1 Univariate post-processing

For each initialization time, the corresponding ensemble forecasts with varying forecast horizons are calibrated using the EMOS methods introduced in Chapter 1. The model parameters are estimated by minimizing the mean CRPS of the predictive distribution, using locally selected training data within a rolling training window (for details about local calibration, refer to Section 3.2). To prevent numerical issues during the minimization of the loss function, local models require extended training periods, and we adopt the optimal training period lengths identified by Hemri *et al.* (2014).

Temperature

The general notation $d = 1, \dots, D$ introduced in Chapter 2 to reference a certain dimension of ensemble forecasts, is now replaced by ℓ , where ℓ -day ahead ECMWF T2M ensemble forecasts ($\ell = 1, 2, \dots, 10$) are calibrated using the normal EMOS model (see Section 1.2.1) with a training period of 720 days. The verification is performed over the period from January 1, 2004 to March 20,

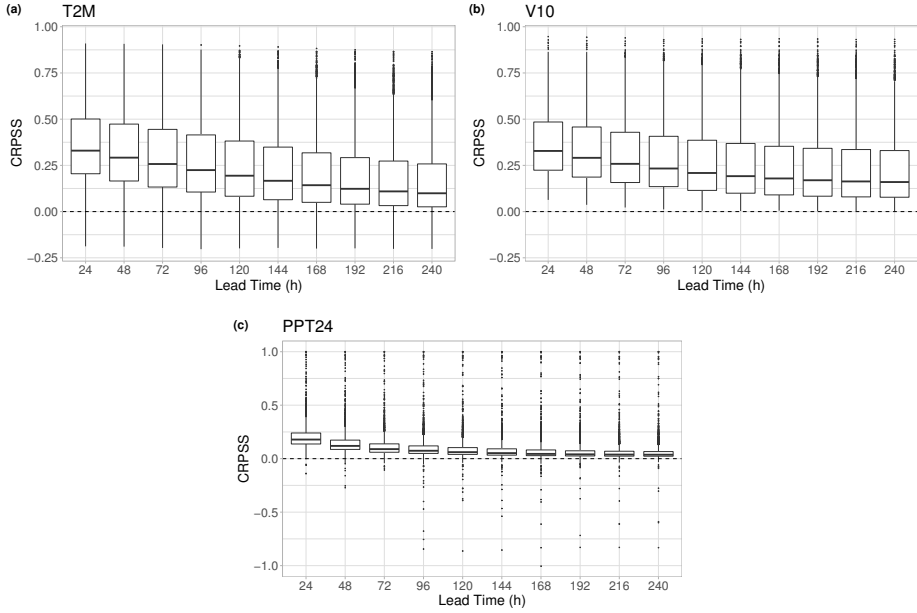


Figure 6.1: Boxplots of the CRPSS of the EMOS models for T2M (a), V10 (b) and PPT24 (c) over the verification period with respect to the corresponding raw ensemble forecasts.

2014, covering 3732 calendar days. Figure 6.1a provides boxplots of the CRPSS for the EMOS models during the verification period, benchmarked against the raw ensemble forecasts.

Compared with the raw ECMWF T2M forecasts, univariate post-processing substantially decreases the mean CRPS for the vast majority of stations for all lead times; however, the gain decreases with the increase of the forecast horizon (see also Feldmann *et al.*, 2019).

Wind speed

The optimal training period length for the truncated normal EMOS model (Section 1.2.2) for post-processing ECMWF V10 ensemble forecasts is 365 calendar days. This allows an almost one-year longer verification period than in the case of T2M; however, to keep the results consistent, we verify both models on the same time interval starting at January 1, 2004. The forecast skill of the EMOS models for various lead times in terms of the mean CRPS over the verification period is depicted in Figure 6.1b. Compared to the T2M ensemble forecasts, EMOS post-processing results in a positive skill score for all consid-

ered stations and the improvement in the mean CRPS is a bit higher, especially for longer lead times.

Precipitation

Precipitation records usually contain a large number of zero observations, hence, reliable parameter estimation in EMOS modelling requires much longer training periods than for temperature or wind speed. Following Hemri *et al.* (2014), to calibrate ECMWF PPT24 forecasts with lead times of $1, 2, \dots, 10$ days, we make use of the EMOS model (Section 1.2.3) with a 1816-day rolling training period leaving the time interval from January 1, 2007 to March 20, 2014 for model verification. Again, in Figure 6.1c we provide the CRPSS values for various forecast horizons, where, similar to the other two weather quantities, the reference forecast is the raw ensemble. The general behaviour of the skill scores is the same as before. For almost all stations EMOS forecasts outperform the raw PPT24 ensemble for all lead times and the improvement decreases with the increase of the forecast horizon. However, compared with T2M and V10, boxplots of the CRPSS values of the EMOS models for PPT24 have much shorter interquartile ranges (IQRs) and display far more outliers.

6.2.2 Multivariate performance

We now continue with the comparison of the forecast skill of the multivariate approaches described in Chapter 2 using calibrated samples of size 52, which is the size of the raw ECMWF ensemble. In methods requiring historical data for providing the dependence template (estimation of the autocorrelation matrix for dECC; observation trajectories for the SSh and for its variants), we consider forecast-observation pairs from the rolling training window applied in univariate calibration. In the case of T2M, in standard SSh variants seasonality is addressed by considering historical data only within 30 calendar days before and after the actual forecast date. For T2M and V10 ensemble forecasts, in mdSSh we select dependence template from observation trajectories of the training data set where at least $m = 6$ from the $L = 10$ observations falls into the corresponding central prediction interval, whereas in the case of PPT24, due to the high computational costs (see also Scheuerer *et al.*, 2017), this method is excluded from the analysis. Additionally, for T2M, tests are also conducted with SSh approaches that incorporate all available historical data, effectively extending the training period. Nevertheless, as will be shown in the next section, these forecasts do not significantly outperform those generated using shorter rolling periods in most cases.

In the following analysis, the ECC-Q forecast is used as a reference for the computation of skill scores and in DM tests for equal predictive performance (see Section 3.1). Further, in the interest of improving the visual presentation

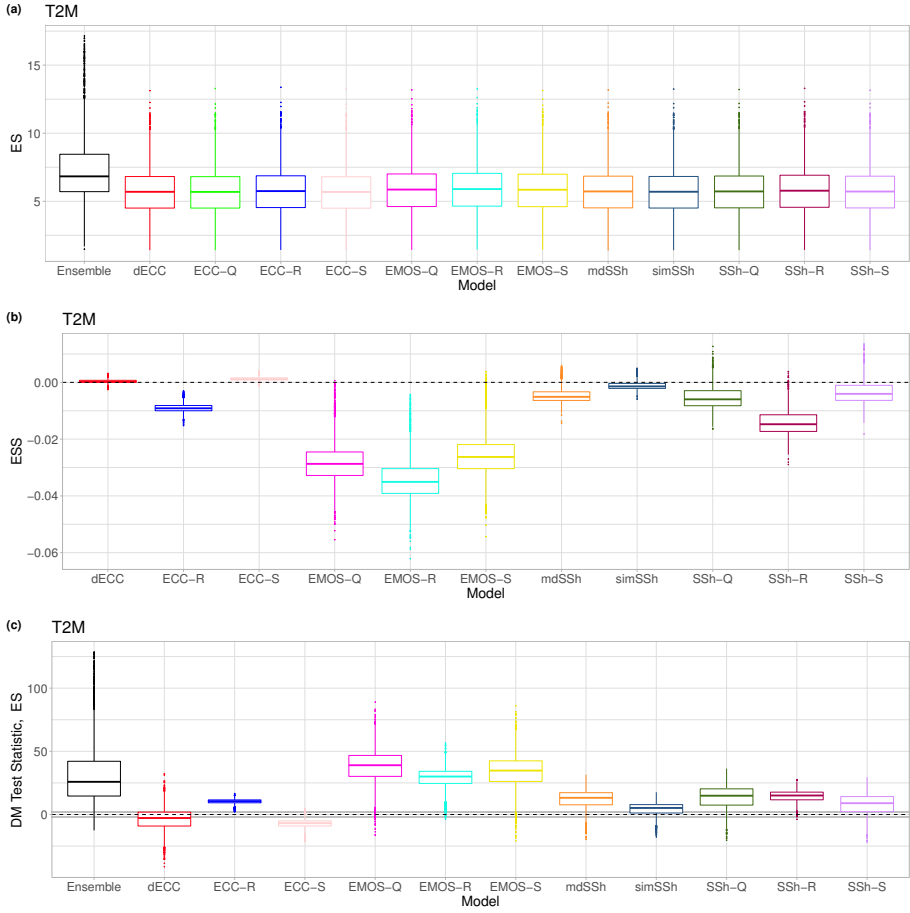


Figure 6.2: Boxplots of the mean ES over the verification period of the calibrated and raw T2M forecasts (a); of the ESS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.

of the results, in the boxplots presented extreme values (further than 3·IQR from the box) are not indicated.

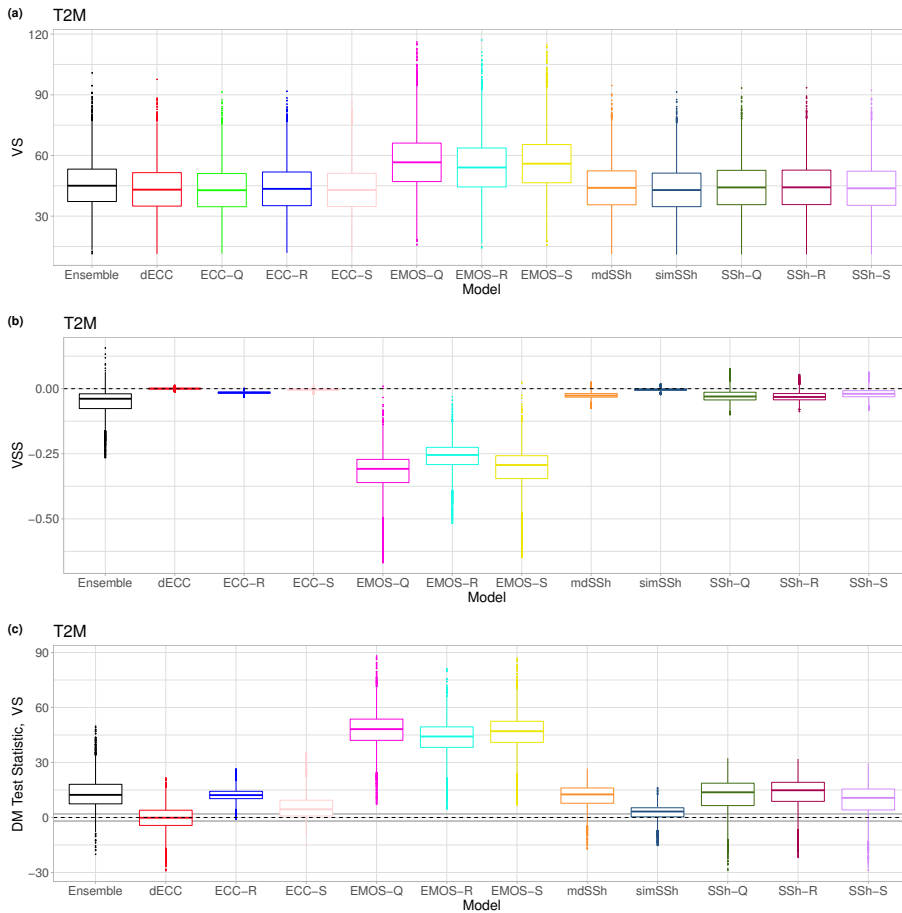


Figure 6.3: Boxplots of the mean VS over the verification period of the calibrated and raw T2M forecasts (a); of the VSS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.

Temperature

In Figure 6.2, the various multivariate post-processing approaches and the raw T2M forecasts are compared in terms of the mean ES (3.1.5) of the different observation stations over the verification period (January 1, 2004 – March 20,

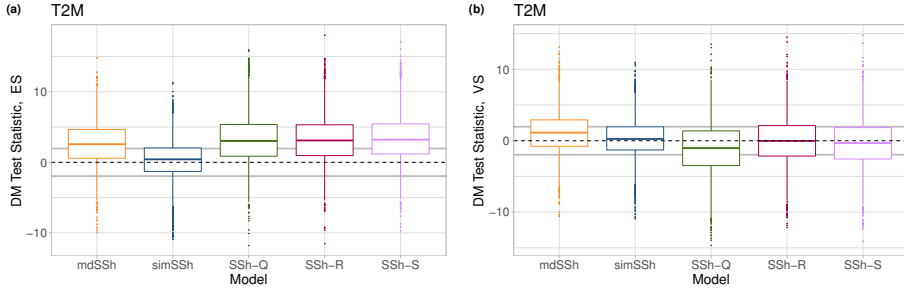


Figure 6.4: Boxplots of the DM test statistic investigating the significance of the difference in the mean ES (a) and mean VS (b) of Schaake shuffle methods based on historical observations from the rolling training period used in EMOS modelling with respect to the corresponding forecasts selecting dependence templates from the whole available past. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.

2014). According to Figure 6.2a, each calibration method improves the ES of the raw ensemble; however, the differences between the various approaches are hardly visible. A better insight can be obtained from Figure 6.2b displaying the ESS values with respect to the reference ECC-Q forecast. The raw ensemble forecasts resulting in very low skill scores are excluded here. All post-processing methods accounting for multivariate dependencies provide clear improvements in terms of the ES over the independent EMOS models, ECC-S shows the best predictive performance, closely followed by ECC-Q and dECC. Perhaps surprisingly, even the advanced variants of the Schaake shuffle (mdSSh and simSSh) fail to outperform the reference ECC-Q method and SSH-S is not behind mdSSh in skill. Finally, according to the results of the DM tests for equal predictive performance (Figure 6.2c), the differences in ES from the reference model are significant for all approaches but dECC.

A similar ranking of the multivariate post-processing methods can be observed in Figure 6.3, though the performance in terms of the VS of the standard SSH variants are closer to each other than in terms of the ES. ECC-Q results in the lowest mean VS; however, its advantage over dECC, ECC-S and simSSh is often not significant (see Figure 6.3c). Further, in contrast to the ES, none of the independent EMOS forecasts outperforms the raw ECMWF ensemble in terms of VS.

As mentioned, SSH approaches with dependence templates selected from the whole available past meaning 721 days for the first and 4443 days for the last day of the verification period have also been tested. As before, in the case of the simple SSH variants historical data only within 30 calendar days around the

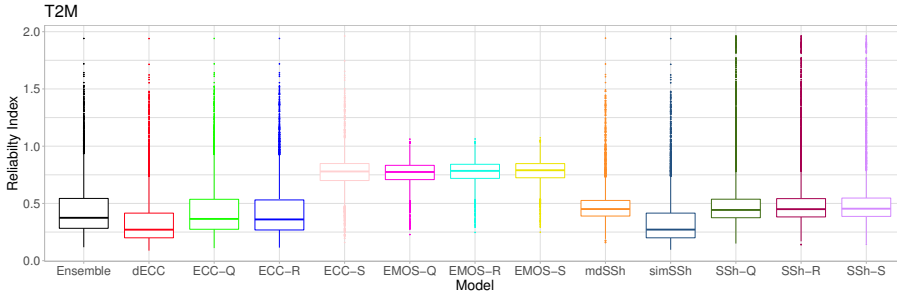


Figure 6.5: Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw T2M forecasts.

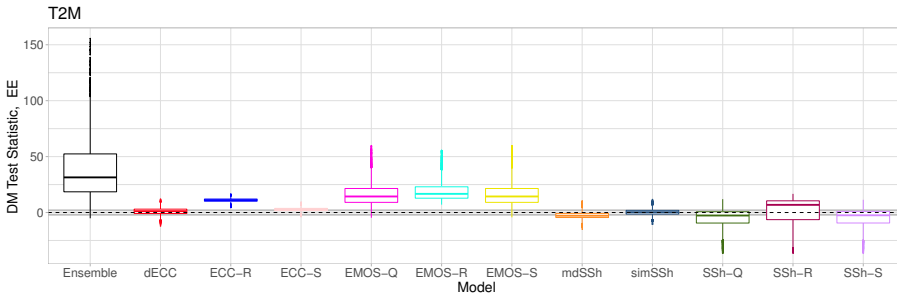


Figure 6.6: Boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors of calibrated and raw T2M forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance.

actual forecast date have been used. However, according to the boxplots of Figure 6.4, the use of the whole available past results in just minor improvements in ES and even smaller in VS, which are not significant. Hence, in the following analysis including case studies with V10 and PPT24 forecasts as well, we proceed with SSH approaches based on historical observations from the rolling training period used in EMOS modelling. This has the potential advantage of accounting more rapidly to substantial changes in the NWP model setup that may influence the forecast error characteristics. For related considerations in the context of univariate post-processing, see Lang *et al.* (2020).

Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw T2M forecasts are displayed in Figure 6.5. ECC-S and all three independent predictions (EMOS-Q, EMOS-

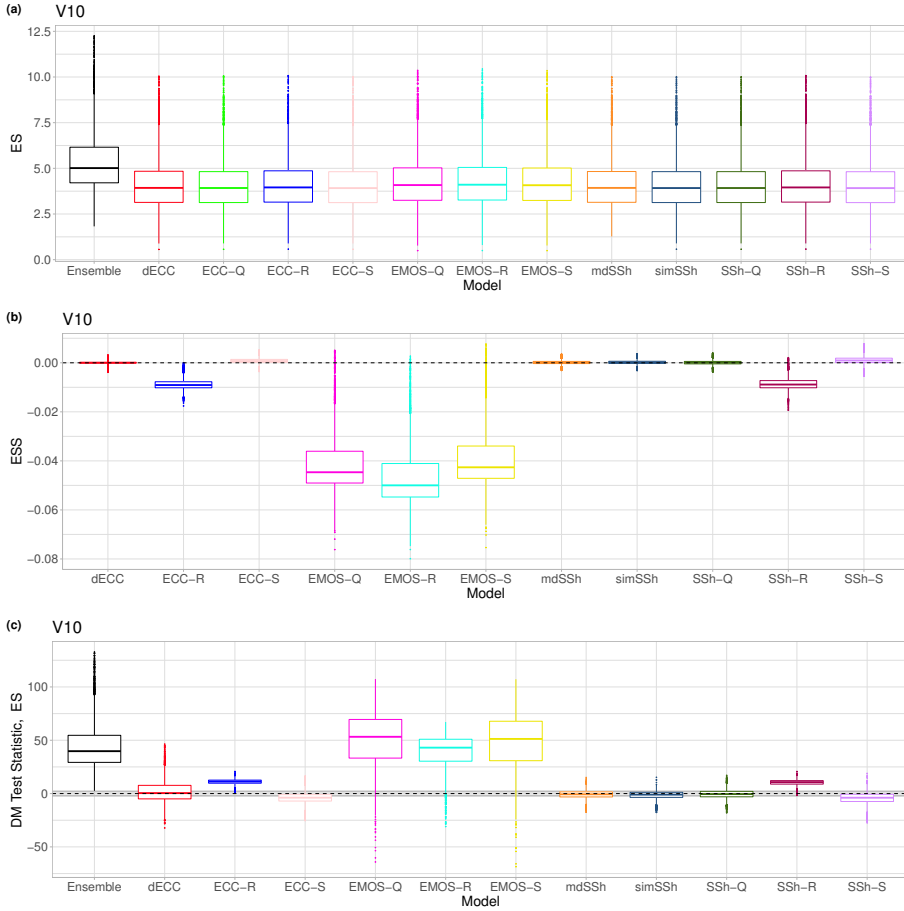


Figure 6.7: Boxplots of the mean ES over the verification period of the calibrated and raw V10 forecasts (a); of the ESS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.

R, EMOS-S), underperform the raw ensemble by a wide margin, which is a consequence of the highly overdispersive character of these forecasts resulting in hump-shaped rank histograms (not shown). SSH-Q, SSH-R and SSH-S still show some overdispersion, whereas the other forecasts, including the raw ensemble, are underdispersive.

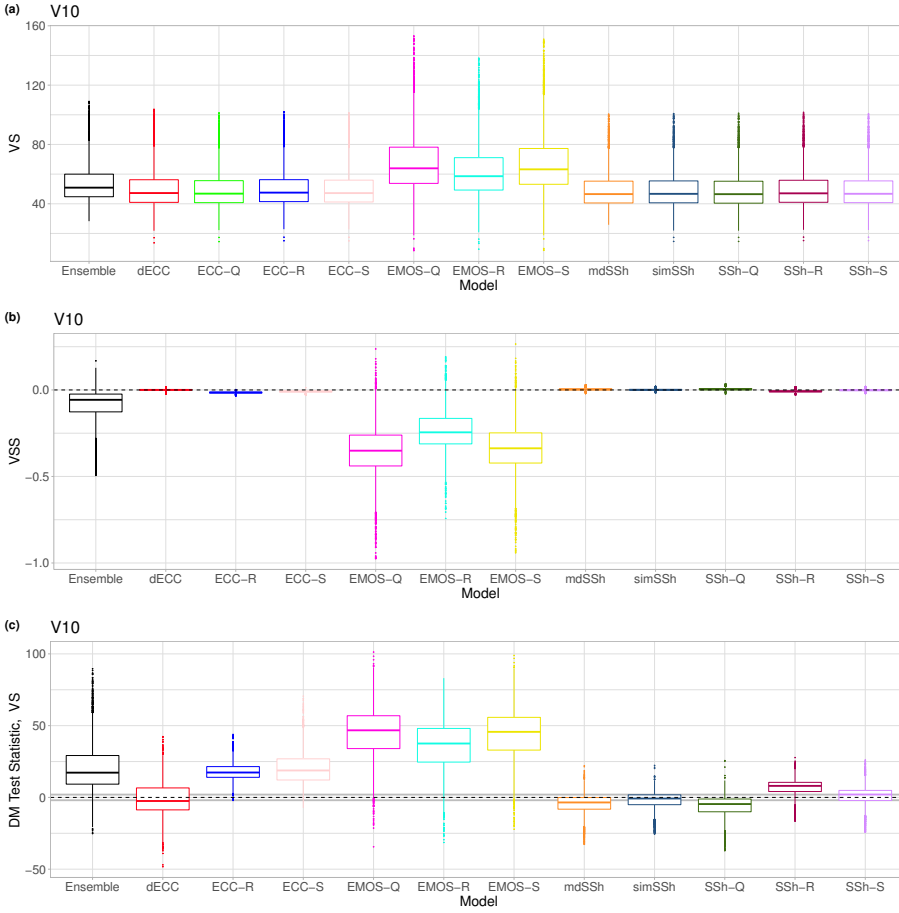


Figure 6.8: Boxplots of the mean VS over the verification period of the calibrated and raw V10 forecasts (a); of the VSS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance.

The average ranks of the dECC and simSSh are the closest to the uniform distribution (the corresponding mean/median reliability indices are 0.357/0.271 for both forecasts), followed by the ECC-R (0.438/0.360) and ECC-Q (0.442/0.365); however, the rank histograms of the latter two forecasts (not shown) are just slightly less underdispersive than that of the ECMWF ensemble

(0.451/0.375), whereas the histogram of the mdSSh is even more U-shaped (0.493/0.451).

Finally, Figure 6.6 shows the boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors. Compared with the raw ensemble, post-processing substantially improves the accuracy of the L^1 median forecast and the empirical copula-based models clearly outperform the independent approaches. The lowest mean EE correspond to the SSh-Q and SSh-S, followed by the mdSSh; however, the differences between these methods and the dECC, ECC-Q, ECC-S and simSSh are not significant.

Note that our ranking of the different post-processing methods differs from results of Heinrich *et al.* (2021) on multivariate calibration of sea surface temperature forecasts, where Schaake shuffle significantly outperforms ensemble copula coupling. However, this disagreement might be explained by the difference in the studied weather quantity and the specific properties of spatial dependencies of gridded sea surface forecasts compared to the temporal dependencies considered here.

Wind speed

The predictive performance of the post-processed and raw V10 ensemble forecast vectors in terms of the mean ensemble ES over the verification period can be investigated with the help of Figure 6.7. Compared with the raw ensemble, each of the investigated post-processing approaches substantially reduces the mean energy score (see Figure 6.7a). According to the skill scores with respect to the ECC-Q model depicted in Figure 6.7b (the raw ensemble is again excluded), the best performing methods are ECC-S and SSh-S, followed by ECC-Q, mdSSh, simSSh and SSh-Q. However, as the DM test statistics provided in Figure 6.7c indicate, the forecast skill of the latter four approaches does not differ significantly. Finally, in contrast to the case of T2M ensemble forecast vectors, all empirical copula-based methods outperform the independent EMOS-Q, EMOS-R and EMOS-S forecasts by a wide margin.

Figure 6.8 summarizing the results for the mean VS is similar to Figure 6.3 in the sense that independent post-processing of raw V10 forecasts with different lead times (EMOS-Q, EMOS-R, EMOS-S) increases the score values. In general, the predictive performance of the empirical copula-based methods is almost identical (see Figure 6.8b), with only ECC-R, ECC-S and SSh-R performing slightly worse. However, according to the results of the DM tests given in Figure 6.8c, the difference in skill of these three methods from the reference ECC-Q approach is significant.

In contrast to temperature, in the case of wind speed any form of post-processing improves the multivariate calibration in the sense that the corresponding average rank histogram is closer to the uniform distribution than the

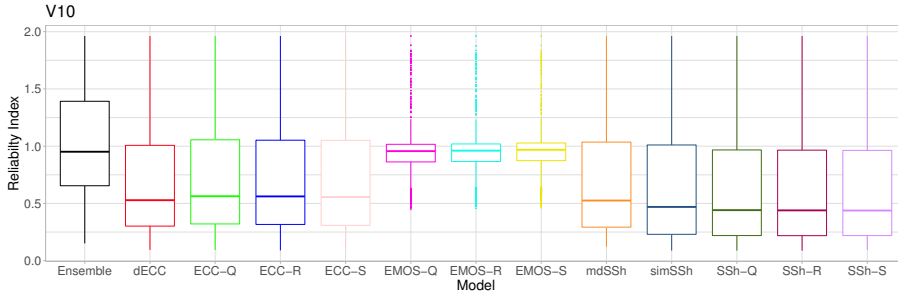


Figure 6.9: Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw V10 forecasts.

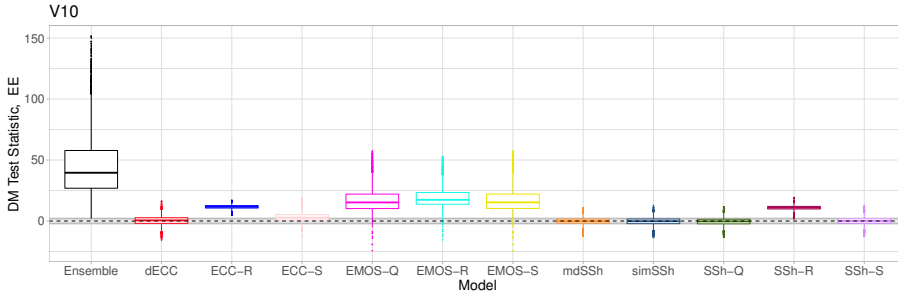


Figure 6.10: Boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors of calibrated and raw V10 forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance.

rank histogram of the raw V10 ensemble. This improvement is quantified in the reliability indices displayed in Figure 6.9, where the highest values again correspond to the independently calibrated EMOS-Q, EMOS-R and EMOS-S forecasts. These approaches result in hump-shaped rank histograms (not shown) indicating overdispersion, the corresponding mean/median RI values are 0.936/0.957, 0.939/0.961 and 0.947/0.968, respectively. All other forecasts (including the raw ensemble) are underdispersive, where the least U-shaped rank histograms correspond to the three versions of the standard Schaake shuffle, followed by the simSSh. The mean/median reliability indices of these forecasts are 0.639/0.441 (SSh-Q), 0.637/0.439 (SSh-R), 0.637/0.438 (SSh-S) and 0.671/0.470 (simSSh), which are substantially better than the corresponding scores of 1.029/0.952 of the raw ensemble vector.

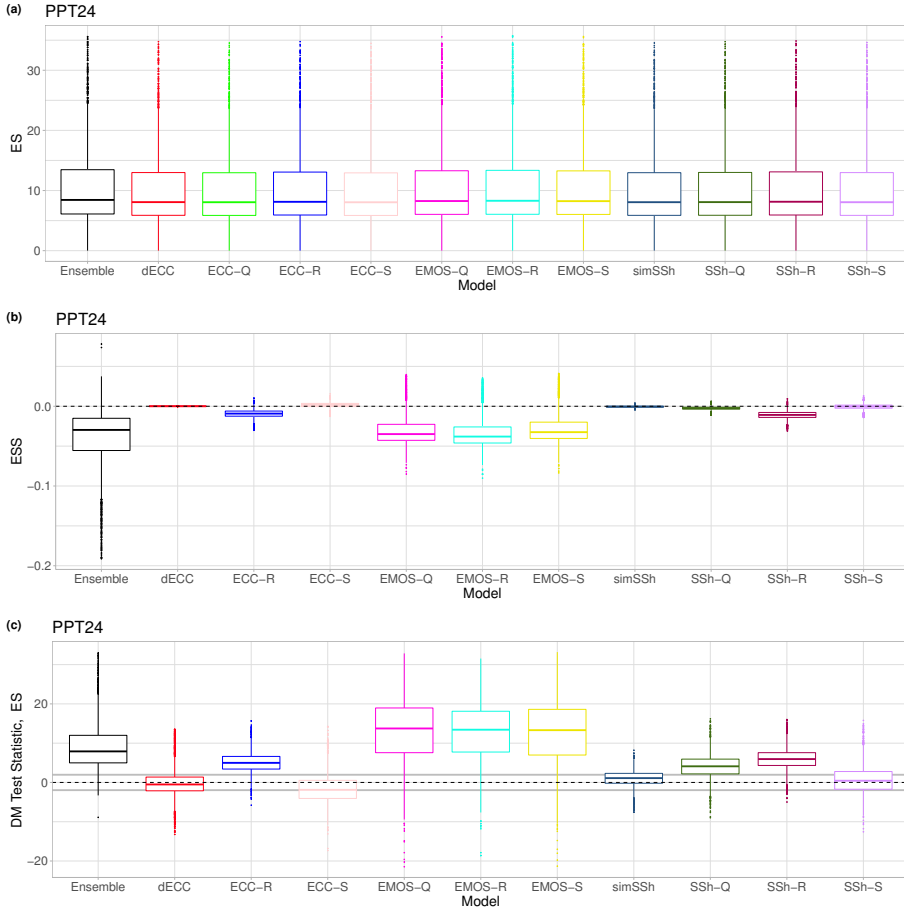


Figure 6.11: Boxplots of the mean ES over the verification period of the calibrated and raw PPT24 forecasts (a); of the ESS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.

Finally, Figure 6.10 showing boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors is almost identical to Figure 6.6. All post-processing methods result in more accurate median forecasts than the raw ensemble and all empirical copula-based methods but ECC-R and SSH-R

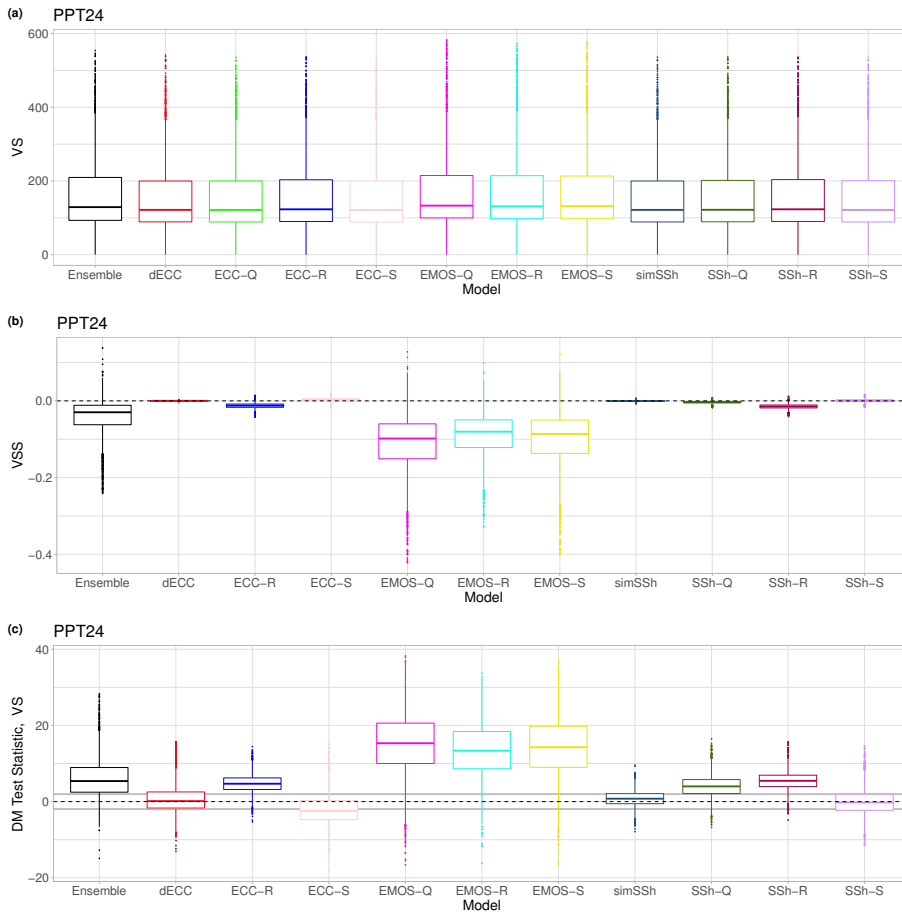


Figure 6.12: Boxplots of the mean VS over the verification period of the calibrated and raw PPT24 forecasts (a); of the VSS with respect to the ECC-Q approach (b); of the DM test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.

perform almost identically well.

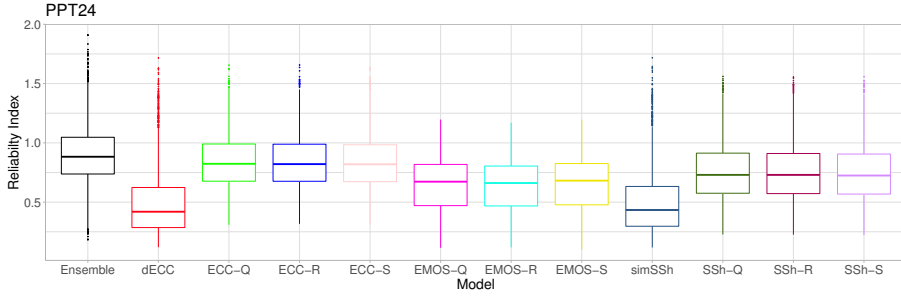


Figure 6.13: Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw PPT24 forecasts.

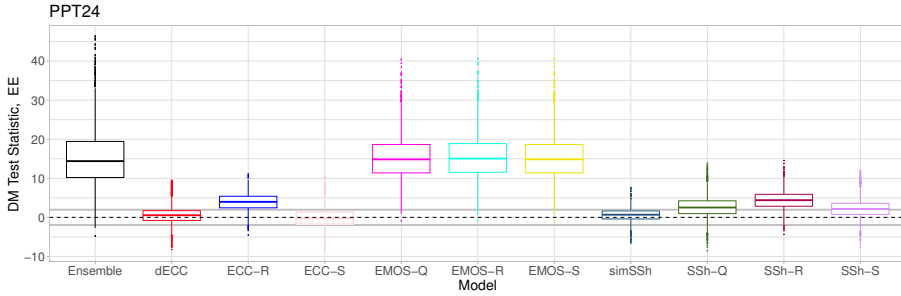


Figure 6.14: Boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors of calibrated and raw PPT24 forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.

Precipitation accumulation

Figure 6.11a displaying the boxplots of the mean ES over the verification period of the calibrated and raw PPT24 forecasts does not reveal a clearly visible difference between the various predictions. The raw ensemble and the independent EMOS-Q, EMOS-R and EMOS-S methods seem to be slightly behind the other forecasts, which is confirmed by the skill scores of Figure 6.11b. From the empirical copula-based approaches ECC-S again shows the best forecast skill, followed by ECC-Q, dECC, simSSH and SSH-S. However, as indicated by the values of the DM test statistics summarized in Figure 6.11c, the differences between the latter four forecasts in terms of the ES are not significant.

In contrast to the other two weather quantities, Figures 6.11 and 6.12 result

in the same conclusions, that is the rankings of the forecasts in terms of the ES and the VS are identical. Further, investigating the boxplots of Figure 6.13, one can again observe the positive effect of post-processing on multivariate calibration quantified in lower reliability indices. The average ranks of dECC and simSSh fit most the uniform distribution, followed by the independent EMOS-Q, EMOS-R and EMOS-S approaches. Note that despite their overdispersive character resulting in hump-shaped rank histograms (not shown), these independent forecasts outperform all versions of the standard SSh and ECC, which are highly underdispersive.

Finally, Figure 6.14 showing the boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors suggests the same ranking of the various post-processing approaches as Figures 6.11 and 6.12. However, in contrast to the ES and VS, with regard to this score even the independent EMOS-Q, EMOS-R and EMOS-S methods outperform the raw ensemble.

6.3 Discussion

Following the emphasis on univariate post-processing methods in the previous chapters, this chapter transitions to a focus on multivariate approaches. We compare a diverse array of state-of-the-art methods for multivariate ensemble post-processing, specifically looking at dependencies over lead times ranging from 1 to 10 days. The findings reveal that these multivariate post-processing methods consistently improve forecast quality compared to both the raw ensemble predictions and straightforward applications of univariate models that overlook multivariate dependencies. The basic ECC and SSh variants showed only slight differences in predictive performance across the different sampling strategies. Specifically, random sampling (ECC-R and SSh-R) tended to perform less effectively than quantile-based (ECC-Q and SSh-Q) or stratified sampling (ECC-S and SSh-S), with no significant distinctions observed between the latter two. In our comparison of the more advanced methods—dECC, mdSSh, and simSSh—against their simpler counterparts, we generally found that these complex techniques did not provide any clear advantages, and there were no instances where they outperformed ECC-Q in a meaningful way.

Overall, our findings reveal that the predictive performance of the various multivariate post-processing methods shows only minor differences. The widely used ECC-Q approach stands out as a robust benchmark due to its straightforward application and low computational costs, making it a preferred choice for future comparative studies. Our results align with those from Lerch *et al.* (2020), which similarly found no single method consistently superior across all potential misspecifications.

To ensure direct comparability among the methods, we limited all ap-

proaches to a fixed sample size based on the raw ensemble. However, one notable advantage of the SSh variants is their capability to generate post-processed ensemble forecasts of arbitrary size. This flexibility could enhance modelling for extreme events (Lerch *et al.*, 2017) and serves as a promising avenue for future research, such as examining how sample size affects performance in relation to recently proposed weighted multivariate proper scoring rules (Allen *et al.*, 2023).

The case studies presented here highlight several opportunities for further exploration and generalization. While our analysis concentrated on temporal dependencies over lead times of 1 to 10 days, it would be valuable to systematically investigate predictive performance concerning spatial or inter-variable dependencies as well. Additionally, we focused on copula-based two-step approaches to multivariate post-processing. However, alternative methods, such as parametric models for the full joint distribution (Baran and Möller, 2015) and quantile mapping (Whan *et al.*, 2021), have also been proposed and could serve as fruitful areas for further comparison. Another potential avenue could be the use of models that adjust the ensemble members in a single step (collectively called member-by-member postprocessing techniques), while preserving the rank order structure of the raw ensemble (Van Schaeybroeck and Vannitsem, 2015). This approach could enable the generation of calibrated forecasts that are consistent across space, time, or multiple weather variables.

Recent advancements in post-processing have illustrated the advantages of incorporating additional predictors to improve the forecasting performance of univariate methods (Rasp and Lerch, 2018). While these advanced techniques can serve as foundational elements for multivariate post-processing schemes, integrating additional predictor information into the second, copula-based step poses challenges. This highlights the need for developing specialized machine learning approaches tailored for multivariate post-processing. Preliminary studies have begun to address this by exploring spatially coherent forecast fields through generative adversarial networks (Dai and Hemri, 2021) and utilizing scoring rule-based generative models for multivariate post-processing that incorporate additional predictors (Chen *et al.*, 2024).

Lastly, evaluating multivariate predictive performance remains a complex issue, as different evaluation metrics can yield varying rankings for the methods considered. Disentangling the contributions to multivariate forecast performance—such as univariate performance, multivariate dependencies, and the quantification of forecast uncertainty—presents significant challenges, especially in understanding how these factors influence the variability observed in metrics like the energy score (ES) and variogram score (VS) in our case studies. Although there has been progress in the methodological aspects of multivariate evaluation (Ziel and Berk, 2019; Alexander *et al.*, 2024; Allen *et al.*, 2023, 2024), there is a pressing need for systematic comparisons of the discrimination

capabilities of multivariate proper scoring rules, ideally utilizing standardized benchmark datasets. As noted by Lerch *et al.* (2020), studies based on large datasets, like the one examined here, could provide valuable insights in this area.

Chapter 7

Summary

Ensemble forecasts play a crucial role in modern meteorology, as they allow for the estimation of uncertainties in model outputs. The ensemble method involves running multiple simulations with slightly perturbed initial conditions, each designed to capture the inherent variability of atmospheric dynamics. The resulting forecasts offer a range of possible outcomes, helping to assess the probabilities of various future weather scenarios (Wilks, 2019, Chapter 8). Accurate weather forecasting is vital across many sectors, including agriculture, transportation, and disaster management. Reliable forecasts support effective decision-making, enhance public safety, and improve resource allocation.

Although ensemble forecasts provide valuable information about weather uncertainties, they have certain drawbacks. One major concern is that the reliability of these forecasts largely depends on the quality of the initial conditions and the models used. If the initial conditions are not accurately defined, or if the model fails to capture key atmospheric processes, the forecasts may be inaccurate (Buizza, 2018b).

Calibration techniques, such as ensemble model output statistics (EMOS; Gneiting *et al.*, 2005), are essential for improving the reliability of ensemble forecasts. EMOS is a statistical model that links ensemble forecasts to actual observations, allowing meteorologists to identify and correct biases, thereby increasing the accuracy of the forecasts. The parameters of the predictive distributions are determined by minimizing an appropriate scoring rule—most commonly the continuous ranked probability score (CRPS; Wilks, 2019, Section 9.5.1)—using historical ensemble forecasts and observations. Due to the characteristics of meteorological variables, it is necessary to use different EMOS models with distinct distributions for calibration. For example, the normal distribution is suitable for modeling temperature, while a censored generalized extreme value distribution is more appropriate for modelling precipitation accumulation.

During univariate post-processing, the model typically focuses on adjusting or calibrating the forecast of each variable independently. Therefore, since it assumes the independence of marginal distributions, multivariate post-processing is crucial for improving the accuracy of weather forecasts by accounting for interactions and dependencies between multiple variables, such as temperature, humidity, and precipitation, as well as between different lead times or stations. Effectively modelling these interactions enhances overall predictive performance and provides a more coherent representation of atmospheric conditions, leading to more reliable forecasts for various applications.

This dissertation, which is based on the author’s four papers, focuses on advancing statistical post-processing methods for weather forecasts and provides a comprehensive comparison of these approaches. We highlight both univariate and multivariate methods to improve the reliability and applicability of forecasts.

In the first paper (Baran and Lakatos, 2024), summarized in Chapter 4, we introduced a novel approach for interpolating parametric predictive distributions to locations where no validation observations for parameter estimation are available, bypassing the computational burdens of existing spatial interpolation methods. To this end, 75% of the 674 available synoptic observation stations were used for parameter estimation and to derive the predictive distributions. In the next step, these distributions were assigned to the remaining 168 stations, which had no observation data, based on feature vectors obtained from the forecasts. This method uses data essential for EMOS modelling and demonstrates that local and observation-based semi-local methods perform better at observed locations, while forecast-based semi-local methods are more effective at unobserved locations, resulting in sharper prediction intervals and improved accuracy metrics such as mean CRPS and mean absolute error. In our case study, which assessed the effectiveness of clustering-based interpolation models, we utilized the 52-member ensemble forecasts of 10-meter wind speed from the European Centre for Medium-Range Weather Forecasts (ECMWF), spanning Europe, Asia, and Australia, for the period from January 1, 2020, to April 30, 2023. Our findings emphasize that no one-size-fits-all solution exists; instead, the choice of post-processing method should be guided by the specific context and data availability.

Chapter 5 synthesizes findings from two previous studies on visibility forecasts (Baran and Lakatos, 2023; Lakatos and Baran, 2024), revealing that both climatological (ensemble forecast determined based on historical observations) and post-processed forecasts significantly outperform raw ensemble predictions for the two datasets used in our papers. First, we analyzed visibility data from the *EUPPBench* benchmark dataset (Demaeyer *et al.*, 2023), which contains the ECMWF control forecast, 50 interchangeable ensemble members generated using random perturbations, and also includes high-resolution forecasts.

According to the recommendations of the World Meteorological Organization, visibility is reported in discrete values ranging from 0 to 70 km, defining a set of 84 elements. As a result, post-processing simplifies to a classification task. In our case studies, we apply some of the models also used by Baran *et al.* (2021), which were employed to calibrate total cloud cover forecasts due to the discrete nature of the observable values. Covering the years 2017 and 2018, our study employed proportional odds logistic regression (POLR; McCullagh, 1980) and multilayer perceptron (MLP; Goodfellow *et al.*, 2016) classifiers, demonstrating notable performance improvements. The second dataset includes operational 51-member ECMWF visibility ensemble forecasts for calendar years 2020-2021, and our findings suggest, that calibrated univariate forecasts achieve a notable improvement compared to raw and climatological forecasts, particularly with the inclusion of Copernicus Atmospheric Monitoring Service (CAMS) visibility point forecasts as additional covariates. Besides, the earlier mentioned classification techniques, we also evaluated the performance enhancements in the simplest multivariate models after incorporating CAMS. We examine two versions of the two-step empirical copula-based post-processing methods: the simplest ensemble copula coupling (ECC; Schefzik *et al.*, 2013) and the Schaake shuffle (SSh; Clark *et al.*, 2004) approaches. These techniques rearrange the samples taken from the calibrated distributions based on the rank order structure of the raw ensemble and historical observations, thereby determining the dependence template of the copula function. Our findings indicate that incorporating CAMS improves energy scores (Gneiting and Raftery, 2007) by 0.58–4.11% compared to forecasts based solely on ECMWF predictions. For variogram score (Scheuerer and Hamill, 2015), CAMS offers approximately a 5% advantage for independent methods, although MLP-C underperforms compared to climatology. While ECC and SSh methods show less than 2.5% improvement relative to ECMWF-based models, they outperform climatology by more than 20%.

Transitioning to the fourth paper (Lakatos *et al.*, 2023), summarized in Chapter 6, we shifted our focus to a comprehensive comparison of state-of-the-art multivariate ensemble post-processing methods, using global ECMWF ensemble forecasts of 2m temperature, 10m wind speed, and 24-hour precipitation accumulation from January 1, 2002, to March 20, 2014. By analyzing dependencies over lead times of 1 to 10 days, we found that these methods consistently improve forecast quality compared to both raw ensemble predictions and simpler univariate models. Among the basic ECC and SSh variants, only minor differences in predictive performance were observed, with the ECC-Q (based on equidistant quantiles of the univariate predictive distributions) approach emerging as a robust benchmark due to its straightforward application and low computational costs. This finding aligns with previous studies that also highlighted the lack of a universally superior method across different scenarios,

reaffirming the need for careful method selection based on specific forecasting contexts.

In conclusion, the findings of this dissertation could not only address important questions in the field of weather forecasting but could also pave the way for numerous future research directions. The growing popularity of machine learning models, particularly convolutional neural networks (Veldkamp *et al.*, 2021; Li *et al.*, 2022), presents significant opportunities for effectively modelling interactions across various dimensions. Moreover, the emergence of EMOS-based models that redefine correlation structures offers exciting possibilities for generating more reliable forecasts. To fully harness these advancements in practical applications, ongoing interdisciplinary collaboration will be essential, ultimately fostering a deeper understanding of atmospheric dynamics.

Chapter 8

Összefoglalás

Az ensemble előrejelzések kulcsfontosságú szerepet töltenek be a modern meteorológiában, mivel lehetővé teszik az előrejelző modellek kimeneteiben rejlő bizonytalanság becslését. Az ensemble módszer lényege, hogy kissé eltérő kezdeti feltételekkel egyszerre több szimulációt futtatnak, amelyek célja, hogy megragadják a légkör dinamikájában rejlő változékonyságot. Az így kapott előrejelzések a lehetséges jövőbeli kimenetek egy széles spektrumát kínálják, segítve ezzel a különböző időjárási forgatókönyvek valószínűségének pontosabb becslését (Wilks, 2019, Chapter 8). A megbízható időjárás-előrejelzések az élet számos területén elengedhetetlenek, például a mezőgazdaságban, közlekedésben és katasztrófavédelemben. A pontos előrejelzések segítik a hatékony döntéshozatalt, növelik a közbiztonságot, és hozzájárulnak az erőforrások jobb elosztásához.

Bár az ensemble előrejelzések értékes információval szolgálnak az időjárás bizonytalanságairól, vannak bizonyos hátrányaik is. Az egyik fő probléma, hogy megbízhatóságuk erősen függ a kezdeti feltételek és a használt modellek minőségétől. Ha a kezdeti feltételek nem pontosak, vagy ha a modell nem képes megfelelően megragadni a légköri folyamatok kulcsfontosságú elemeit, akkor az előrejelzések pontatlanok lehetnek (Buizza, 2018b).

Az utófeldolgozó eljárások, mint például az ensemble model output statistics (EMOS; Gneiting *et al.*, 2005), elengedhetetlenek az ensemble előrejelzések megbízhatóságának javításához. Az EMOS egy statisztikai modell, amely összekapcsolja az ensemble előrejelzéseket a valós megfigyelésekkel, lehetővé téve a meteorológusok számára, hogy azonosítsák és korrigálják a torzításokat, ezzel növelve az előrejelzések pontosságát. Az előrejelző eloszlások paraméterei egy megfelelő illeszkedési mutató – leggyakrabban a continuous ranked probability score (CRPS; Wilks, 2019, Section 9.5.1) – minimalizálásával állnak elő, amelyhez múltbeli ensemble előrejelzések és megfigyelések szolgálnak alapul. Az előrejelzések kalibrálására az időjárási mennyiségek természetéből adódóan

különböző eloszlású EMOS modellek alkalmazása indokolt. Míg a hőmérséklet modellezéséhez a normál eloszlás, a csapadékmennyiség esetében a cenzorált általánosított extrémérték-eloszlás lehet megfelelő választás.

Az egyváltozós utófeldolgozás során a modell tipikusan arra fókuszál, hogy egymástól függetlenül javítsa az egyes változók előrejelzéseit. Ennélfogva, mivel ez a technika a marginális eloszlások függetlenségére épít, a többváltozós utófeldolgozás kulcsfontosságú az időjárás-előrejelzések pontosságának növelésében. Ez utóbbi ugyanis figyelembe veszi a különböző változók, például a hőmérséklet, páratartalom és csapadék, illetve a különböző előrejelzési horizontok vagy helyszínek közötti interakciókat és függőségeket. Ezen összefüggések hatékony modellezése javítja az előrejelzési képességet, és pontosabb képet ad a légkör állapotáról, így megbízhatóbb előrejelzéseket biztosít különböző alkalmazások számára.

Jelen disszertáció a szerző négy korábbi munkájára épül és az időjárás-előrejelzések statisztikai utófeldolgozására szolgáló módszerek fejlesztésére, valamint átfogó összehasonlítására fókuszál. Benne kitérünk mind az egyváltozós, mind pedig a többváltozós eljárásokra annak érdekében, hogy javítsuk az előrejelzések megbízhatóságát és alkalmazhatóságát.

Az első tárgyalt cikkünkben (Baran and Lakatos, 2024), amelyet a 4. fejezetben foglalunk össze, egy új megközelítést mutattunk be paraméteres előrejelző eloszlások interpolálására olyan helyszínekre, ahol nem rendelkezünk a paraméterbecsléshez szükséges validáló megfigyelésekkel, amely elkerüli a meglévő térbeli interpolációs módszerek magas számítási költségeit. Ebből a célból a rendelkezésre álló 674 szinoptikus megfigyelőállomás 75%-át használtuk fel paraméterbecslésre, illetve az előrejelző eloszlások meghatározásához, amelyeket a következő lépésben az előrejelzésekből meghatározott jellemző vektorok alapján rendeltük hozzá a fennmaradt 168, megfigyelésekkel nem rendelkező állomáshoz.

Az általunk javasolt módszer kizárólag az EMOS modellezéshez szükséges adatokon alapul és igazolja, hogy azokban a pontokban, ahol vannak megfigyelések a lokális, valamint a megfigyeléseket is használó klaszterezés-alapú szemi-lokális módszerek teljesítenek jobban, míg a megfigyelésekkel nem rendelkező helyszíneken a kizárólag az előrejelzéseken alapuló szemi-lokális technikák a hatékonyabbak. Az új módszer alkalmazásával szűkebb előrejelzési intervallumokat és jobb illeszkedés mutatókat, például kisebb átlagos CRPS értéket és alacsonyabb átlagos abszolút eltérést érhetünk el. A klaszterezés-alapú interpolációs modellek hatékonyságát vizsgáló esettanulmányunkhoz a Középtávú Időjárás Előrejelzések Európai Központjának (ECMWF) Európa, Ázsia és Ausztrália összesen 674 szinoptikus megfigyelőállomásra vonatkozó, a 2020. január 1. és 2023. április 30. közötti időszakot felölelő 52-tagú 10 méteres ensemble előrejelzéseit használtuk. A kapott eredmények arra is rávilágítanak, hogy nincs egyetlen univerzális megoldás; az utófeldolgozó módszert leginkább

a konkrét helyzet és az elérhető adatok figyelembevételével kell kiválasztani.

A 5. fejezet két korábbi tanulmány (Baran and Lakatos, 2023; Lakatos and Baran, 2024) látástávolság előrejelzésekkel kapcsolatos eredményeit összegezi. Bemutatja, hogy a vizsgálatokhoz használt mindkét adatállomány esetén mind a klimatológiai (múltbeli megfigyelések alapján meghatározott ensemble előrejelzés), mind pedig az utófeldolgozott előrejelzések jelentősen felülmúlják a nyers ensemble előrejelzéseket. Először az *EUPPBench* benchmark adatállományból származó látástávolság előrejelzéseket elemeztük (Demaeyer *et al.*, 2023), amely tartalmazza az ECMWF kontroll előrejelzéseket, az 50, véletlenszerű perturbációkkal generált felcserélhető ensemble tagokat, valamint a magas felbontású determinisztikus előrejelzéseket is. A Meteorológiai Világszervezet ajánlása szerint a látástávolságot 0-tól 70 km-ig terjedő diszkrét értékekben jelentik, amely egy 84 elemű halmazt határoz meg. Ennek következtében az utófeldolgozás egy osztályozási feladatra egyszerűsödik. Esettanulmányaink során a Baran *et al.* (2021) által javasolt modellek egy részét alkalmazzuk a látástávolság előrejelzések kalibrálására, amelyeket a megfigyelt értékek diszkrét természetéből adódóan hasonló módon használtak fel a felhősödés előrejelzések kalibrálására. Így, a 2017-es és 2018-as évek adatain alapuló kutatásunkban az utófeldolgozáshoz a rendezett logisztikus regressziót (POLR; McCullagh, 1980) és többrétegű perceptron (MLP; Goodfellow *et al.*, 2016) osztályozót használtunk, amik jelentős teljesítményjavulást eredményeztek. A második vizsgált adatállomány az operatív 51 tagú ECMWF látástávolság ensemble előrejelzéseket tartalmazza a 2020-2021 időszakra. Itt az eredményeink arra utalnak, hogy a kalibrált egyváltozós előrejelzések jelentős javulást mutatnak a nyers és klimatológiai előrejelzésekhez képest, különösen, amikor a Copernicus Atmospheric Monitoring Service (CAMS) kategorikus előrejelzéseit is bevonjuk a modellek magyarázó változói közé. Ezen túlmenően, az említett osztályozási technikák mellett azt is elemeztük, mennyit javul a legegyszerűbb többváltozós modellek teljesítménye a CAMS előrejelzések bevonása után. Ehhez a kétlépcsős empirikus kopula-alapú utófeldolgozó módszerek két változatát vizsgáljuk: a legegyszerűbb (ECC; Schefzik *et al.*, 2013) és a Schaake shuffle (SSh; Clark *et al.*, 2004) módszereket. Ezen technikák az egyváltozós utófeldolgozás alkalmazása után a nyers ensemble és a múltbeli megfigyelések rangstruktúrája alapján rendezik át a kalibrált eloszlásokból vett mintákat, ezáltal meghatározva a kopula függvény függőségi sablonját. Eredményeink azt mutatják, hogy a csupán az ECMWF előrejelzéseken alapuló előrejelzésekhez képest a CAMS használata 0,58–4,11%-kal javítja az energy score (Gneiting and Raftery, 2007) mutatót. A variogram score (Scheuerer and Hamill, 2015) esetében a CAMS körülbelül 5%-os előnyt kínál az független utófeldolgozott előrejelzésekből alkotott vektorok esetén, bár a szemi-lokális MLP-C a klimatológiai előrejelzésekkel összevetve aluteljesít. Míg az ECC és SSh módszerek kevesebb, mint 2,5%-os javulást mutatnak a pusztán ECMWF alapú modellekkel szemben, a klima-

tológiával összehasonlítva már több mint 20%-os előnyt jelentenek.

A negyedik cikkünkben ((Lakatos *et al.*, 2023)), amelyet a 6. fejezetben foglalunk össze, az ECMWF globális ensemble előrejelzéseit felhasználva áttértünk a legmodernebb többváltozós ensemble utófeldolgozó módszerek átfogó összehasonlítására. Itt a 2002. január 1. és 2014. március 20. közötti időszak 2 méteres hőmérséklet, 10 méteres szélsősebesség és 24 órás csapadékösszeg előrejelzéseket vizsgáltuk. A 1–10 napos előrejelzési horizontok közötti függőségek elemzésével arra jutottunk, hogy a nyers ensemble előrejelzésekkel és az egyszerűbb egyváltozós modellekkel összehasonlítva, ezek a módszerek következetesen javítják a predikciók minőségét. A legegyszerűbb ECC és SSh variánsok előrejelző képessége között csak kisebb különbségek voltak megfigyelhetők, ahol egyszerű alkalmazhatósága és alacsony számítási költsége révén (az egydimenziós előrejelző eloszlások ekvidisztáns kvantilisein alapuló) az ECC-Q megközelítés erőteljes referenciaként emelkedett ki. Ez az eredmény összhangban van a korábbi kutatásokkal, amelyek szintén hangsúlyozták, hogy az összes lehetséges forgatókönyvet áttekintve nem létezik univerzálisan legjobb módszer, megerősítve az alkalmazandó technika gondos kiválasztásának a szükségességét, melynek a konkrét előrejelzési problémához kell igazodnia.

Összegzésül, a disszertáció eredményei nemcsak az időjárás-előrejelzés területén felmerülő fontos kérdések megválaszolásában nyújthatnak segítséget, hanem számos jövőbeli kutatási irányvonalat is kijelölhetnek. A gépi tanulási modellek, különösen a konvolúciós neurális hálózatok (Veldkamp *et al.*, 2021; Li *et al.*, 2022) egyre növekvő népszerűsége jelentős lehetőségeket kínál a különböző dimenziók közötti interakciók hatékony modellezésére. Ezen túlmenően az EMOS-alapú modellek megjelenése, amelyek újra definiálják a korrelációs struktúrákat, izgalmas lehetőségeket kínálnak a megbízhatóbb előrejelzések előállítására. Ahhoz, hogy ezeket az előnyöket a gyakorlati alkalmazásokban teljes mértékben kiaknázhassuk, elengedhetetlen a folytatódó interdiszciplináris együttműködés, amely végső soron hozzájárul a légkör dinamikájának mélyebb megértéséhez.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Sándor Baran, for his invaluable guidance and the opportunities he has provided to foster my academic development. His inspiring teaching, which began with my very first Statistics class during my undergraduate studies, ignited my passion for education.

Most importantly, I would like to express my deepest love and gratitude to my husband, Dávid, and our family, whose unwavering belief in me and constant motivation have been my greatest sources of strength. Your support has made this journey possible, and I am forever grateful for your love and encouragement.

I also acknowledge that the studies included in this dissertation were made with the support of the ÚNKP-23-3 and ÚNKP-22-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund and of the National Research, Development and Innovation Office under Grant No. K142849 and the S&T cooperation program 2021-1.2.4-TÉT-2021-00020.

“As an act of bravery, love cannot be sentimental; as an act of freedom, it must not serve as a pretext for manipulation. It must generate other acts of freedom; otherwise, it is not love.”

— Paulo Freire, *Pedagogy of the Oppressed*

Bibliography

- Alexander, C., Coulon, M., Han, Y. and Meng, X. (2024) Evaluating the discrimination ability of proper multi-variate scoring rules. *Ann. Oper. Res.* **334**, 857–883.
- Allen, S., Ginsbourger, D. and Ziegel, J. (2023) Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA J. Uncertain. Quantif.* **11**, 906–940.
- Allen, S., Ziegel, J. and Ginsbourger, D. (2024) Assessing the calibration of multivariate probabilistic forecasts. *Q. J. R. Meteorol. Soc.* **150**, 1315–1335.
- Baran, Á. and Baran, S. (2024) A two-step machine learning approach to statistical post-processing of weather forecasts for power generation. *Q. J. R. Meteorol. Soc.* **150**, 1029–1047.
- Baran, Á., Lerch, S., El Ayari, M. and Baran, S. (2021) Machine learning for total cloud cover prediction. *Neural. Comput. Appl.* **33**, 2605–2620.
- Baran, S., Baran, Á., Pappenberger, F. and Ben Bouallègue, Z. (2020) Statistical post-processing of heat index ensemble forecasts: is there a royal road? *Q. J. R. Meteorol. Soc.* **146**, 3416–3434.
- Baran, S. and **Lakatos, M.** (2023) Statistical post-processing of visibility ensemble forecasts. *Meteorol. Appl.* **30**, paper e2157, doi:10.1002/met.2157.
- Baran, S. and **Lakatos, M.** (2024) Clustering-based spatial interpolation of parametric post-processing models. *Wea. Forecasting* **9**, 1591–1604.
- Baran, S. and Lerch, S. (2015) Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Q. J. R. Meteorol. Soc.* **141**, 2289–2299.
- Baran, S., Leutbecher, M., Szabó, M. and Ben Bouallègue, Z. (2019) Statistical post-processing of dual-resolution ensemble forecasts. *Q. J. R. Meteorol. Soc.* **145**, 1705–1720.

- Baran, S. and Möller, A. (2015) Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics* **26**, 120–132.
- Baran, S., Szokol, P. and Szabó, M. (2021) Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts. *Environmetrics* **32**, paper e2678, doi:10.1002/env.2678.
- Baran, S. and Nemoda, D. (2016) Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics* **27**, 280–292.
- Bauer, P., Thorpe, A. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55.
- Ben Bouallègue, Z., Haiden, T., Weber, N. J., Hamill, T. M. and Richardson, D. S. (2020) Accounting for representativeness in the verification of ensemble precipitation forecasts. *Mon. Weather Rev.* **148**, 2049–2062.
- Ben Bouallègue, Z., Heppelmann, T., Theis, S. E. and Pinson, P. (2016) Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Mon. Weather Rev.* **144**, 4737–4750.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300.
- Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X. and Tian, Q. (2023) Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538.
- Boudala, F. S., Isaac, G. A., Crawford, R. W. and Reid, J. (2012) Parameterization of runway visual range as a function of visibility: Implications for numerical weather prediction models. *J. Atmosphere Ocean. Technol.* **29**, 177–191.
- Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B. and Beare, S. E. (2008) The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **134**, 703–722.
- Buizza, R. (2018a) Introduction to the special issue on “25 years of ensemble forecasting”. *Q. J. R. Meteorol. Soc.* **145**, 1–11.
- Buizza, R. (2018b) Ensemble forecasting and the need for calibration. In Vanitsem, S., Wilks, D. S., Messner, J. W. (eds.), *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, Amsterdam, pp. 15–48.

- Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* **133**, 1076–1097.
- Bremnes, J. B. (2020) Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Mon. Weather Rev.* **148**, 403–414.
- Breiman, L. (2001) Random forests. *Mach. Learn.* **45**, 5–32.
- Chen, J., Janke, T., Steinke, F. and Lerch, S. (2024) Generative machine learning methods for multivariate ensemble post-processing. *Ann. Appl. Stat.* **18**, 159–183.
- Chmielecki, R. M. and Raftery, A. E. (2011) Probabilistic visibility forecasting using Bayesian model averaging. *Mon. Weather Rev.* **139**, 1626–1636.
- Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R. (2004) The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**, 243–262.
- Dabernig, M., Mayr, G. J., Messner, J. W. and Zeileis, A. (2017) Spatial ensemble post-processing with standardized anomalies. *Q. J. R. Meteorol. Soc.* **143**, 909–916.
- Dai, Y. and Hemri, S. (2021) Spatially coherent postprocessing of cloud cover ensemble forecasts. *Mon. Weather Rev.* **149**, 3923–3937.
- Dawid and Sebastiani (1999) Coherent dispersion criteria for optimal experimental design. *Ann. Stat.* **27**, 65–81.
- Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X. and Stull, R. B. (2006) Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.* **111**, paper D24307, doi:10.1029/2005JD006917.
- Demaeyer, J., Bhend, J., Lerch, S., Primo, C., Van Schaeybroeck, B., Atencia, A., Ben Bouallègue, Z., Chen, J., Dabernig, M., Evans, G., Faganeli Pucer, J., Hooper, B., Horat, N., Jobst, D., Merše, J., Mlakar, P., Möller, A., Mestre, O., Taillardat, M. and Vannitsem, S. (2023) The EUPPBench postprocessing benchmark dataset v1.0. *Earth Syst. Sci. Data* **15**, 2635–2653.
- Diebold, F. X. and Mariano, R. S. (1995) Comparing predictive accuracy. *J. Bus. Econ. Stat.* **13**, 253–263.
- Dietz, S. J., Kneringer, P., Mayr, G. J. and Zeileis, A. (2019) Low-visibility forecasts for different flight planning horizons using tree-based boosting models. *Adv. Stat. Clim. Meteorol. Oceanogr.* **5**, 101–114.

- ECMWF (2024a) *IFS Documentation CY49R1 – Part IV Physical processes*. ECMWF, Reading. Available at: <http://dx.doi.org/10.21957/c731ee1102> [Accessed on 6 January 2025]
- ECMWF (2024b) *IFS Documentation CY49R1 – Part V: Ensemble Prediction System*. ECMWF, Reading, UK, Available at: <http://dx.doi.org/10.21957/956d60ad81>. [Accessed on 6 January 2025]
- ECMWF Directorate (2012) Describing ECMWF’s forecasts and forecasting system. *ECMWF Newsletter* **133**, 11–13.
- Epstein, E. S. (1969) Stochastic dynamic prediction. *Tellus*, **21**, 739–759.
- Feldmann, K., Richardson, D. S. and Gneiting, T. (2019) Grid- versus station-based postprocessing of ensemble temperature forecasts. *Geophys. Res. Lett.* **46**, 7744–7751.
- Friedman, J. H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232.
- Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–411.
- Gneiting, T. (2011) Making and evaluating point forecasts. *J. Am. Stat. Assoc.* **106**, 746–762.
- Gneiting, T. (2014) Calibration of medium-range weather forecasts. *ECMWF Technical Memorandum* No. 719. Available at: <http://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf> [Accessed on 25 November 2024].
- Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007) Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Series B Stat. Methodol.* **69**, 243–268.
- Gneiting, T. and Raftery, A. E. (2005) Weather forecasting with ensemble methods. *Science* **310**, 248–249.
- Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction and estimation. *J. Amer. Statist. Assoc.* **102**, 359–378.
- Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**, 1098–1118.

- Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econ. Stat.* **29**, 411–422.
- Gneiting, T., Stanberry, L. I., Gneiting, 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**, 211–235.
- Good, I. J. (1952) Rational decisions. *J. R. Stat. Soc. Series B Stat. Methodol.* **14**, 107–114.
- Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. MIT Press, Cambridge.
- Gultepe, I., Müller, M. D. and Boybeyi, Z. (2006) A new visibility parameterization for warm-fog applications in numerical weather prediction models. *J. Appl. Meteorol. Climatol.* **45**, 1469–1480.
- Haiden, T., Janousek, M., Vitart, F., Ben Bouallègue, Z., Ferranti, L., Prates, F. and Richardson, D. (2021) Evaluation of ECMWF forecasts, including the 2021 upgrade. *ECMWF Technical Memorandum* No. 884. Available at: <http://dx.doi.org/10.21957/90pgicjk4> [Accessed on 25 November 2024]
- Haslett, J. and Raftery, A. E. (1989) Space–time modelling with long-memory dependence: assessing Ireland’s wind power resource (with discussion) *J. R. Stat. Soc. Ser. C Appl. Statist.* **38**, 1–50.
- He, D., Wang, Y., Tang, Y., Kong, D., Yang, J., Zhou, W., Li, H., Wang, F. (2024) Improvement in the forecasting of low visibility over Guizhou, China, based on a multi-variable deep learning model. *Atmosphere* **15**, paper 752, doi:10.3390/atmos15070752.
- Heinrich, C., Hellton, K. H., Lenkoski, A. and Thorarinsdottir, T. L. (2021) Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. *J. Amer. Statist. Assoc.* **116**, 1048–1059.
- Hemri, S., Haiden, T. and Pappenberger, F. (2016) Discrete post-processing of total cloud cover ensemble forecasts. *Mon. Weather Rev.* **144**, 2565–2577.
- Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. and Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.* **41**, 9197–9205.
- Hu, Y., Schmeits, M. J., van Andel, J. S., Verkade, J. S., Xu, M., Solomatine, D. P. and Liang, Z. (2016) A stratified sampling approach for improved sampling from a calibrated ensemble forecast distribution. *J. Hydrometeorol.* **17**, 2405–2417.

- Izenman, A. J. (2008) *Modern Multivariate Statistical Techniques. Regression, Classification and Manifold Learning*. Springer, New York.
- Jávoriné Radnóczy, K., Várkonyi, A. and Szépszó, G. (2020) On the way towards the AROME nowcasting system in Hungary. *ALADIN-HIRLAM Newsletter* **14**, 65–69.
- Jobst, D., Möller, A. and Groß, J. (2023) D-vine-copula-based postprocessing of wind speed ensemble forecasts. *Q. J. R. Meteorol. Soc.* **149**, 2575–2597.
- Jordan, A., Krüger, F. and Lerch, S. (2019) Evaluating probabilistic forecasts with scoringRules. *J. Stat. Softw.* **90**, 1–37.
- Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F. and Gritmit, E. (2011) Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Weather Rev.* **139**, 2630–2649.
- Knüppel, M. (2015) Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *J. Bus. Econ. Stat.* **33**, 270–281.
- Knüppel, M., Krüger, F. and Pohle, M.-O. (2022) Score-based calibration testing for multivariate forecast distributions. *Preprint*. Available at: <https://doi.org/10.48550/arXiv.2211.16362>. [Accessed on 25 November 2024]
- Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K. and Anandkumar, A. (2023) FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference (PASC '23)*, (Article No. 13, 11 pages) Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3592979.3593412>.
- Lakatos, M.** and Baran S. (2024) Enhancing multivariate post-processed visibility predictions utilizing Copernicus Atmosphere Monitoring Service forecasts. *Meteorol. Appl.* **31**, paper e70015, doi:10.1002/met.70015
- Lakatos, M.**, Lerch, S., Hemri, S. and Baran, S. (2023) Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts. *Q. J. R. Meteorol. Soc.* **149**, 856–877.
- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Meroze, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S. and Battaglia, P. (2023) Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421.

- Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Ben Bouallègue, Z., Clare, M., Lessig, C., Magnusson, L., Prieto Nemesio, A. (2024) AIFS: a new ECMWF forecasting system. *ECMWF Newsletter* **178**, 4–5.
- Lang, M. N., Lerch, S., Mayr, G. J., Simon, T., Stauffer, R. and Zeileis, A. (2020) Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Process. Geophys.* **27**, 23–34.
- Leith, C. E. (1974) Theoretical skill of Monte-Carlo forecasts. *Mon. Weather Rev.* **102**, 409–418.
- Lerch, S. and Baran, S. (2017) Similarity-based semi-local estimation of EMOS models. *J. R. Stat. Soc. Ser. C Appl. Statist.* **66**, 29–51.
- Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S. and Graeter, M. (2020) Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Process. Geophys.* **27**, 349–371.
- Lerch, S. and Thorarinsdottir, T. L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A* **65**, 21206.
- Lerch, S., Thorarinsdottir, T.L., Ravazzolo, F. and Gneiting, T. (2020) Forecaster’s dilemma: Extreme events and forecast evaluation. *Stat. Sci.* **32**, 106–127.
- Li, W., Pan, B., Xia, J. and Duan, Q. (2022) Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *J. Hydrol.* **605**, paper 127301, doi:10.1016/j.jhydrol.2021.127301.
- Lorenz, E. N. (1963) Deterministic nonperiodic flow. *J. Atmosphere Sci.* **20**, 130–141.
- López-Pintado, S. and Romo, J. (2009) On the concept of depth for functional data. *J. Amer. Statist. Assoc.* **104**, 718–734.
- Marzban, C., Leyton, S. and Colman, B. (2007) Ceiling and visibility forecasts via neural networks. *Wea. Forecasting* **22**, 466–479.
- McCullagh, P. (1980) Regression model for ordinal data (with discussion) *J. R. Stat. Soc. Series B Stat. Methodol.* **42**, 243–268.
- Möller, A. and Groß, J. (2020) Probabilistic temperature forecasting with a heteroscedastic autoregressive ensemble postprocessing model. *Q. J. R. Meteorol. Soc.* **146**, 211–224.

- Möller, A., Lenkoski, A. and Thorarinsdottir, T. L. (2013) Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Q. J. R. Meteorol. Soc.* **139**, 982–991.
- Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996) The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.
- Pahlavan, R., Moradi, M., Tajbakhsh, S., Azadi, M. and Rahnama, M. (2021) Fog probabilistic forecasting using an ensemble prediction system at six airports in Iran for 10 fog events. *Meteorol. Appl.* **28**, paper e2033, doi:10.1002/met.2033.
- Parde, A. N., Ghude, S. D., Dhangar, N. G., Lonkar, P., Wagh, S., Govardhan, G., Biswas, M. and Jenamani, R. K. (2022) Operational probabilistic fog prediction based on ensemble forecast system: A decision support system for fog. *Atmosphere* **13**, paper 1608, doi:10.3390/atmos13101608.
- Penov, N. and Guerova, G. (2023) Sofia Airport visibility estimation with two machine-learning techniques. *Remote Sens.* **15**, paper 4799, doi:10.3390/rs15194799.
- Perrone, E., Schicker, I. and Lang, M. N. (2020) A case study of empirical copula methods for the statistical correction of forecasts of the ALADIN-LAEF system. *Meteorol. Z.*, **29**, 277–288.
- Pillosu, F. M., Hewson, T. D., Stephens, E., Prudhomme, C. and Cloke, H. L. (2024) Does a multivariate approach enhance univariate grid-to-point post-processed rainfall forecasts? A comparative analysis. *ECMWF Technical Memorandum* **924**
- Politis, D. N. and Romano, J. P. (1994) The stationary bootstrap. *J. Amer. Stat. Assoc.*, **89**, 1303–1313.
- Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R. and Willson, M. (2025) Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90.
- Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**, 1155–1174.
- Rasp, S. and Lerch, S. (2018) Neural Networks for Postprocessing Ensemble Weather Forecasts. *Mon. Weather Rev.* **164**, 3885–3900.

- Rüschendorf, L. (2009) On the distributional transform, Sklar's theorem, and the empirical copula process. *J. Stat. Plann. Inference* **139**, 3921–3927.
- Roquelaure, S. and Bergot, T. (2009) Contributions from a local ensemble prediction system (LEPS) for improving fog and low cloud forecasts at airports. *Wea. Forecasting* **24**, 39–52.
- Ryerson, W. R. and Hacker, J. P. (2018) A nonparametric ensemble postprocessing approach for short-range visibility predictions in data-sparse areas. *Wea. Forecasting* **33**, 835–855.
- Schefzik, R. (2016) Combining parametric low-dimensional ensemble postprocessing with reordering methods. *Q. J. R. Meteorol. Soc.* **142**, 2463–2477.
- Schefzik, R. and Möller, A. (2018) Ensemble postprocessing methods incorporating dependence structures. In Vannitsem, S., Wilks, D. S. and Messner, J. W. (eds.), *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, Amsterdam, pp. 91–125.
- Schefzik, R., Thorarinsdottir T. L. and Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28**, 616–640.
- Scheuerer, M. (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.* **140**, 1086–1096.
- Scheuerer, M. and Büermann, L. (2014) Spatially adaptive post-processing of ensemble forecasts for temperature. *J. R. Stat. Soc. Ser. C Appl. Statist.* **63**, 405–422.
- Scheuerer, M. and Hamill, T. M. (2015) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Mon. Weather Rev.* **143**, 1321–1334.
- Scheuerer, M., Switanek, M. B., Worsnop, R. P. and Hamill, T. M. (2020) Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Weather Rev.* **148**, 3489–3506.
- Scheuerer, M., Hamill, T. M., Whitin, B., He, M. and Henkel, A. (2017) A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.* **53**, 3029–3046.
- Scheuerer, M. and Möller, D. (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.* **9**, 1328–1349.

- Schulz, B., El Ayari, M., Lerch, S. and Baran, S. (2021) Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting. *Sol. Energy* **220**, 1016–1031.
- Schultz, B. and Lerch, S. (2022) Machine learning methods for postprocessing ensemble forecasts of wind gusts: a systematic comparison. *Mon. Weather Rev.* **150**, 235–257.
- Shankar, A. and Sahana, B. C. (2023) Early warning of low visibility using the ensembling of machine learning approaches for aviation services at Jay Prakash Narayan International (JPNI) Airport Patna. *SN Appl. Sci.* **5**, paper 132, doi:10.1007/s42452-023-05350-7.
- Sklar, A. (1959) Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8**, 229–231.
- Szabó, M., Gascón, E. and Baran, S. (2023) Parametric post-processing of dual-resolution precipitation forecasts. *Wea. Forecasting* **38**, 1313–1322.
- Taillardat, M. (2021) Skewed and mixture of Gaussian distributions for ensemble postprocessing. *Atmosphere* **12**, paper 966, doi:10.3390/atmos12080966.
- Taillardat, M. and Mestre, O. (2020) From research to applications – examples of operational ensemble post-processing in France using machine learning. *Nonlin. Processes Geophys.* **27**, 329–347.
- Taillardat, M., Mestre, O., Zamo, M. and Naveau, P. (2016) Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Weather Rev.* **144**, 2375–2393.
- Thorarinsdottir, T. L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *J. R. Stat. Soc. Ser. A Stat. Soc.* **173**, 371–388.
- Thorarinsdottir, T. L., Gneiting, T. and Gissibl, N. (2013) Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertainty Quantif.* **1**, 522–534.
- Thorarinsdottir, T. L., Scheuerer, M. and Heinz, C. (2016) Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Stat.* **25**, 105–122.
- Vardi, Y. and Zhang, C. H. (2000) The multivariate L_1 -median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97**, 1423–1426.
- Van Schaeybroeck, B. and Vannitsem, S. (2015) Ensemble post-processing using member-by-member approaches: theoretical aspects. *Q. J. R. Meteorol. Soc.* **141**, 807–818.

- Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Boualègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Odak Plenković, I., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K. and Ylhaisi, J. (2021) Statistical postprocessing for weather forecasts – review, challenges and avenues in a big data world. *Bull. Amer. Meteorol. Soc.* **102**, E681–E699.
- Veldkamp, S., Whan, K., Dirksen, S. and Schmeits, M. (2021) Statistical post-processing of wind speed forecasts using convolutional neural networks. *Mon. Weather Rev.* **149**, 1141–1152.
- Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S. Fourth Edition*. Springer, New York.
- Wagh, S., Kulkarni, R., Lonkar, P., Parde, A. N., Dhargar, N. G., Govardhan, G., Sajjan, V., Debnath, S., Gultepe, I., Rajeevan, M. and Ghude, S. D. (2023) Development of visibility equation based on fog microphysical observations and its verification using the WRF model. *Model. Earth Syst. Environ.* **9**, 195–211.
- Whan, K., Zscheischler, J., Jordan, A.I. and Ziegel, J.F. (2021) Novel multivariate quantile mapping methods for ensemble post-processing of medium-range forecasts. *Weather. Clim. Extremes* **32**, paper 100310, doi:10.1016/j.wace.2021.100310.
- Wilks, D. S. (2006) Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorol. Appl.* **13**, 243–256.
- Wilks, D. S. (2016) "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Am. Meteorol. Soc.* **97**, 2263–2273.
- Wilks, D. S. (2018) Univariate ensemble postprocessing. In Vannitsem, S., Wilks, D. S., Messner, J. W. (eds.), *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, Amsterdam, pp. 49–89.
- Wilks, D. S. (2019) *Statistical Methods in the Atmospheric Sciences. 4th ed.* Elsevier, Amsterdam.
- Wilks, D. S. (2020) Regularized Dawid–Sebastiani score for multivariate ensemble forecasts. *Q. J. R. Meteorol. Soc.* **146**, 2421–2431.
- World Meteorological Organization (1992) *International Meteorological Vocabulary* (WMO-No.182) WMO, Geneva.

- World Meteorological Organization (2018) *Guide to Instruments and Methods of Observation. Volume I – Measurement of Meteorological Variables* (WMO-No.8) WMO, Geneva.
- World Meteorological Organization (2022) *Aerodrome Reports and Forecasts: A Users' Handbook to the Codes* (WMO-No. 782) WMO, Geneva.
- Zell, A., Mache, N., Hübner, R., Mamier, G., Vogt, M., Schmalzl, M. and Herrmann, K.-U. (1994) SNNS (Stuttgart Neural Network Simulator). In Skrzypek, J. (ed.), *Neural Network Simulation Environments*. The Kluwer International Series in Engineering and Computer Science, vol 254. Springer, Boston.
- Zhou, B., Du, J., Gultepe, I. and Dimego, G. (2012) Forecast of low visibility and fog from NCEP: Current status and efforts. *Pure Appl. Geophys.* **169**, 895–909.
- Zhou, X., Zhu, Y., Hou, D., Fu, B., Li, W., Guan, H., Sinsky, E., Kolczynski, W., Xue, X., Luo, Y., Peng, J., Yang, B., Tallapragada, V. and Pegion, P. (2022) The development of the NCEP global ensemble forecast system version 12. *Wea. Forecasting* **37**, 1069–1084.
- Zhou, B., Du, J., McQueen, J. and Dimego, G. (2009) Ensemble forecast of ceiling, visibility, and fog with NCEP Short-Range Ensemble Forecast system (SREF) *Aviation, Range, and Aerospace Meteorology Special Symposium on Weather–Air Traffic Management Integration*, Phoenix, AZ, American Meteorological Society, extended abstract 4.5. Available at: https://ams.confex.com/ams/89annual/techprogram/paper_142255.htm [Accessed on 25 November 2024]
- Ziel, F. and Berk, K. (2019) Multivariate forecasting evaluation: On sensitive and strictly proper scoring rules. *Preprint*. Available at: <https://doi.org/10.48550/arXiv.1910.07325> [Accessed on 25 November 2024]

Appendix A

List of publications and conferences

Journal papers

Lakatos, M., Lerch, S., Hemri, S. and Baran, S. (2023) Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts. *Q. J. R. Meteorol. Soc.* **149**, 856–877.

<https://doi.org/10.1002/qj.4436>

(**IF: 3; SJR: D1**) [Ind. cit.: 10]

Baran, S., and **Lakatos, M.** (2023) Statistical post-processing of visibility ensemble forecasts. *Meteorol. Appl.* **30**, paper e2157.

<https://doi.org/10.1002/met.2157>

(**IF: 2.3; SJR: Q2**) [Ind. cit.: 2]

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

<https://doi.org/10.1175/WAF-D-24-0016.1>

(**IF: 3; SJR: Q1**)

Lakatos, M. and Baran S. (2024) Enhancing multivariate post-processed visibility predictions utilizing Copernicus Atmosphere Monitoring Service forecasts. *Meteorol. Appl.* **31**, paper e70015.

<https://doi.org/10.1002/met.70015>

(**IF: 2.3; SJR: Q2**)

Unrelated papers

Baran, S., Marín, J. C., Cuevas, O., Díaz, M., Szabó, M., Nicolis, O. and **Lakatos, M.** (2025) Machine learning-based probabilistic forecasting of solar irradiance in Chile. *Adv. Stat. Climatol. Meteorol. Oceanogr.*, in press.

(SJR: Q2)

International conferences

Nagy-Lakatos, M. and Baran, S.: Enhancing multivariate post-processed visibility predictions utilizing CAMS forecasts, *18th International Conference of the ERCIM WG on Computational and Methodological Statistics*, London, UK, 14-16 December 2024 (invited session talk)

Nagy-Lakatos, M. and Baran, S.: Machine learning-based discrete post-processing of visibility ensemble forecasts (poster presentation), *EGU General Assembly 2024*, Vienna, Austria, 14–19 Apr 2024, EGU24-5541, <https://doi.org/10.5194/egusphere-egu24-5541>, 2024.

Nagy-Lakatos, M. and Baran, S.: Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts, *16th International Conference of the ERCIM WG on Computational and Methodological Statistics*, Berlin, Germany, 16-18 December 2023 (invited session talk)

Nagy-Lakatos, M. and Baran, S.: Classification-based calibration of visibility ensemble forecasts, *16th German Probability and Statistics Days*, Essen, Germany, 7-10 March 2023

Nagy-Lakatos, M. and Baran, S.: Statistical post-processing of visibility ensemble forecasts, *15th International Conference of the ERCIM WG on Computational and Methodological Statistics*, London, UK, 17-19 December 2022

Lakatos, M. and Baran, S.: Restoration of temporal dependence in post-processed ensemble forecasts, *EGU General Assembly 2022*, Vienna, Austria, 23–27 May 2022, EGU22-13125, <https://doi.org/10.5194/egusphere-egu22-13125>, 2022

Lakatos, M. and Baran, S.: Restoration of temporal dependence in statistical post-processing of ensemble weather forecasts, *6th conference of the Deutsche Arbeitsgemeinschaft Statistik*, Hamburg, Germany, March 28 - April 1, 2022

Lakatos, M.: Comparison of multivariate ensemble post-processing methods, *25th Young Statisticians Meeting*, Vorau, Austria. 2021 (invited session talk)

Baran, S. and Lakatos, M.: Comparison of multivariate post-processing methods (poster presentation). *The 1st Conference on Information Technology and Data Science*, Debrecen, Hungary, 2020

Appendix B

ECMWF visibility ensemble calibration

As mentioned in Chapter 5, this appendix presents the results of the calibration of the 51-member ECMWF ensemble forecasts for the 2020-2021 period. With the inclusion of CAMS forecasts, comparable forecast horizons were available up to 120 hours with steps of 6 hours, however, this appendix examines the entire 240-hour forecast horizon range.

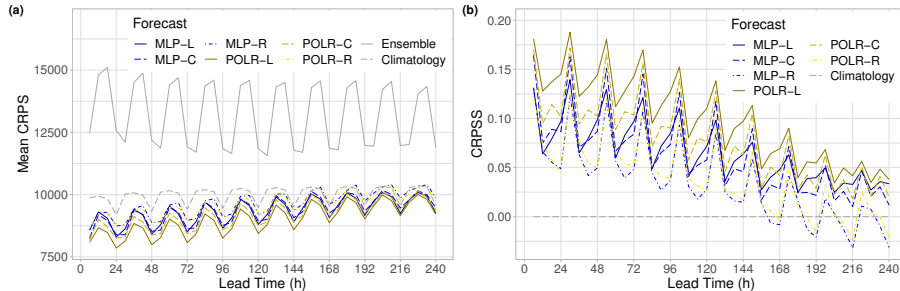


Figure B.1: Mean CRPS of post-processed, raw and climatological visibility forecasts for calendar year 2021 (a) and CRPSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.

According to Figure B.1a, in terms of the mean CRPS all post-processing methods outperform the raw ECMWF ensemble forecast for all forecast horizons by a wide margin and climatology also performs rather well. The clear diurnal cycle in the mean CRPS corresponds to the different observation times

	MLP-L	MLP-C	MLP-R	POLR-L	POLR-C	POLR-R	Climatology
CRPS	70.65 %	70.75 %	72.94 %	68.09 %	69.60 %	72.52 %	75.60 %
LogS	45.31 %	44.10 %	44.88 %	44.28 %	43.45 %	44.44 %	66.05 %

Table B.1: Overall mean CRPS/LogS of post-processed and climatological visibility forecasts for calendar year 2021 as proportion of the mean CRPS/LogS of the raw ECMWF ensemble.

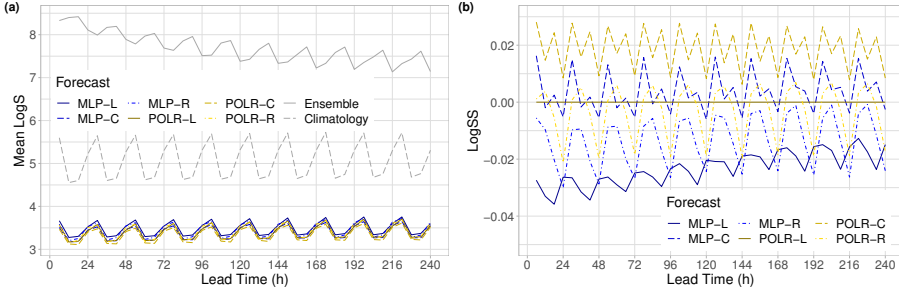


Figure B.2: Mean LogS of post-processed, raw and climatological visibility forecasts for calendar year 2021 (a) and LogSS of post-processed forecasts with respect to the POLR-L model (b) as functions of the lead time.

(0000, 0600, 1200 and 1800 UTC), indicating that the skill of the predictions depends on the time of the day. Raw ensemble forecasts show the best predictive performance at 0600 UTC, whereas both for climatology and for post-processed predictions the lowest mean CRPS is attained at 0000 UTC. Note that the superiority of climatology over the raw visibility forecasts is fully in line with the results of Chmielecki and Raftery (2011). A deeper insight into the differences between the various calibration techniques can be obtained from Figure B.1b depicting the CRPSS values with respect to climatology. In general, clustering-based modelling is superior to regional, and the advantage of post-processing with respect to climatology decreases with the increase of the lead time. For POLR models the chosen training period length is long enough even for reliable local estimation and the POLR-L approach results in the highest skill score for all lead times. This is not the case for the MLP models, where the CRPSS of the semi-locally estimated MLP-C is often above the skill score of the MLP-L, especially for lead times corresponding to 0600 UTC. POLR models outperform their MLP counterparts for most of the forecast horizons; however, up to 156 h even the poorest MLP-R has better skill than climatology. A possible ranking of the competing methods can be obtained from Table B.1 giving the overall mean score values of calibrated and climatological forecasts as proportions of

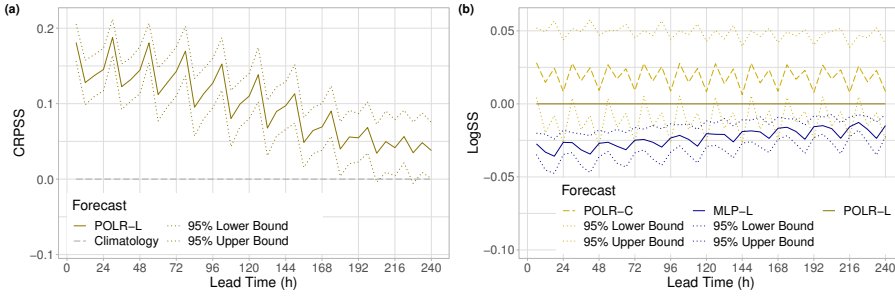


Figure B.3: CRPS of POLR-L post-processed visibility forecasts for calendar year 2021 with respect to climatology (a) and LogSS of POLR-C and MLP-L approaches with respect to the POLR-L model (b) together with 95 % confidence intervals as functions of the lead time.

the corresponding mean score of the raw ECMWF ensemble. POLR-L results in the lowest mean CRPS, closely followed by POLR-C.

As Figure B.2a and Table B.1 show, the difference between raw and post-processed forecasts in terms of the mean LogS is even more pronounced than in terms of the mean CRPS; however, the general picture slightly changes. On the one hand, all post-processing methods outperform climatology by a wide margin. On the other hand, in terms of the LogS, clustering-based modelling is clearly superior to local estimation. According to Figure B.2b, for most of the forecast horizons MLP-L underperforms even MLP-R, and at lead times corresponding to 0600 and 1200 UTC, POLR-R also results in positive LogSS with respect to POLR-L.

To enlighten the uncertainty of the investigated verification scores and statistical significance of the score differences, in Figure B.3 the CRPS and LogSS values of some post-processing approaches are accompanied with 95 % confidence intervals. According to Figure B.3a, the advantage of the best-performing POLR-L model over climatology in terms of the mean CRPS is clearly significant for all lead times but 204 h and 228 h. Regarding the LogS (Figure B.3b), POLR-L significantly outperforms MLP-L for all lead times, whereas the difference between POLR-L and the best performing POLR-C is significant at a 5 % level only at forecast horizons corresponding to 0600 UTC.

Furthermore, PIT histograms of Figure B.4 also demonstrate the positive effect of post-processing and the superiority of climatology over the raw visibility ensemble forecasts. The histograms of the latter are rather U-shaped with a more and more pronounced bias with the increase of the forecast horizon. All other forecasts result in PIT values that are much closer to the desired standard uniform distribution; however, climatology still exhibits a small underdis-

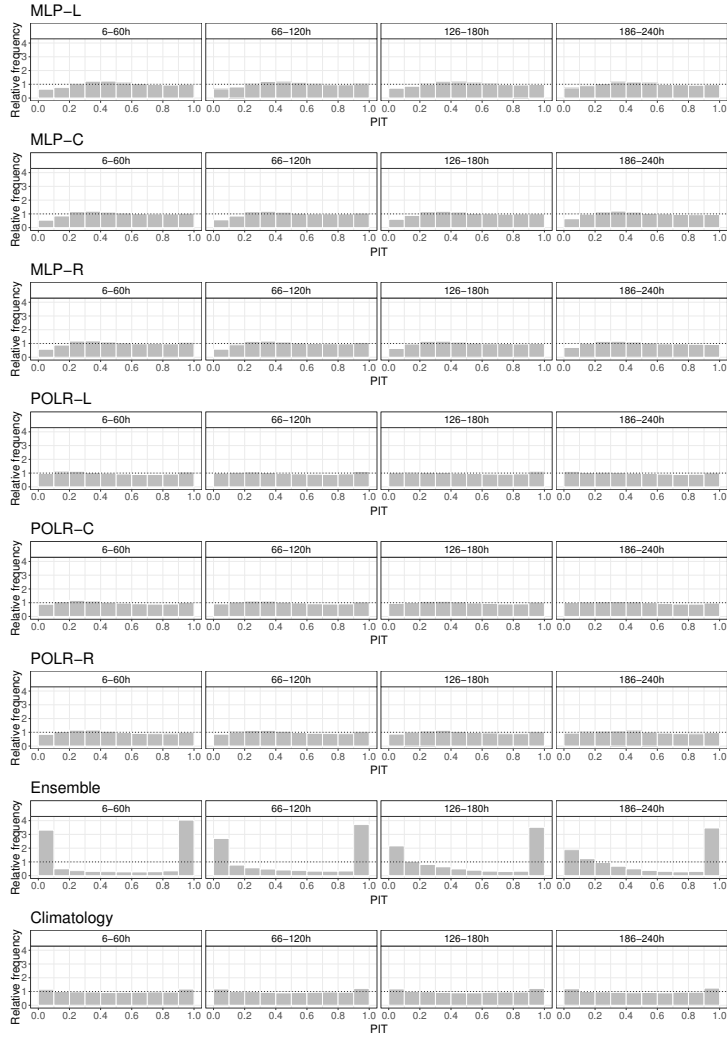


Figure B.4: PIT histograms of post-processed, climatological and raw visibility forecasts for calendar year 2021 for lead times 6–60 h, 66–120 h, 126–180 h and 186–240 h.

person, whereas MLP post-processed predictions are slightly overdispersive. Similar to the case of *EUPPBench* forecasts (Section 5.3.1), the α_{1234}^0 rejects the uniformity for all calibrated forecasts for all lead times; however, now the

MLP-L	MLP-C	MLP-R	POLR-L	POLR-C	POLR-R
1.98×10^{-15}	6.54×10^{-14}	1.15×10^{-7}	4.31×10^{-5}	1.54×10^{-4}	2.00×10^{-5}

Table B.2: Overall mean p -values of the α_{1234}^0 tests for uniformity of the PIT values for calendar year 2021.

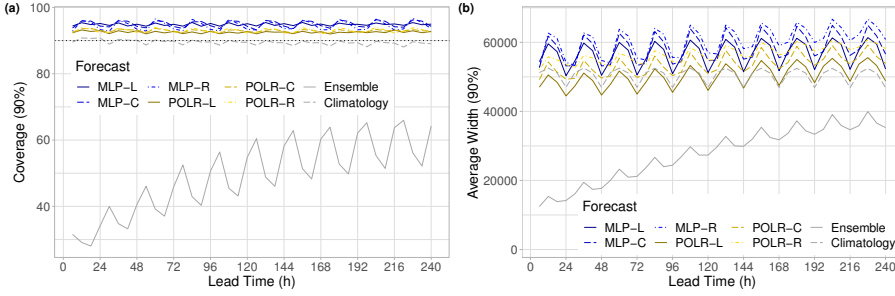


Figure B.5: Coverage (a) and average widths (b) of 90 % central prediction intervals of raw and post-processed visibility forecasts for calendar year 2021 as functions of the lead time. In panel (a), the ideal coverage is indicated by the horizontal dotted line.

semi-local POLR model results in the highest overall mean p -value, see Table B.2.

The improved calibration of post-processed and climatological forecasts can also be observed in Figure B.5a showing the coverage of 90 % central prediction intervals. For the raw forecasts this coverage ranges from 28 % to 66 % and shows an increasing trend with a clear diurnal cycle. For post-processed and climatological forecasts the diurnal cycle is far less pronounced and one cannot observe dependence on the forecast horizon. The coverage of climatology is almost perfect, whereas post-processed forecasts result in coverages slightly above 90 %. For the POLR models the mean absolute deviations over all lead times from the desired 90 % are 2.53 % (POLR-L), 2.96 % (POLR-C) and 2.99 % (POLR-R), while for the MLP approaches one gets 4.86 % (MLP-R), 4.98 % (MLP-C) and 4.52 % (MLP-L). Note that for climatological forecasts this mean absolute deviation is just 0.69 %. Naturally, the price for the better calibration of post-processed and climatological forecasts has to be paid in the loss of the sharpness. As depicted in Figure B.5b, the raw ECMWF visibility ensemble results in far the narrowest 90 % central prediction intervals. This is a rather common phenomenon stemming from the highly underdispersive character of the raw forecasts, which in our case improves with the increase of the lead

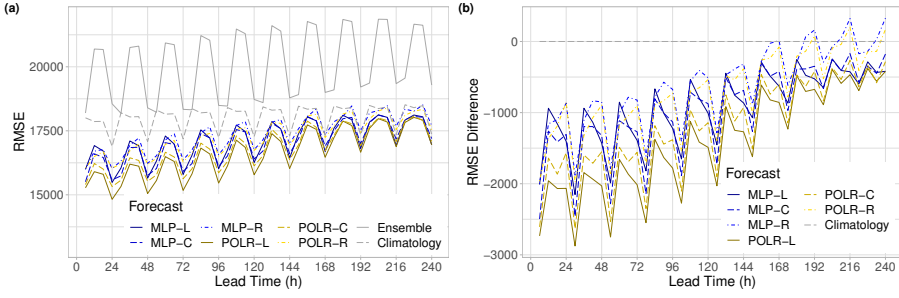


Figure B.6: RMSE of the mean forecasts for calendar year 2021 (a) and difference in RMSE from climatology (b) as functions of the lead time.

time (see Figure B.4). Post-processing usually increases the spread of the predictions, which directly leads to deterioration in sharpness. Furthermore, one should also note that in general, the average width values of the different forecasts are fairly consistent with the corresponding coverages.

Finally, according to Figure B.6a, compared with the raw ensemble all POLR and MLP models substantially improve the accuracy of the mean forecast, whereas the advantage of climatology is far less pronounced, especially for shorter forecast horizons. Note that the ranking of the different forecasts with respect to the RMSE of the mean (Figure B.6b) is completely in line with the ordering based on the mean CRPS (Figure B.1b), and the POLR-L approach again outperforms the competitors for all lead times.

Again, besides the 30-day reference climatology, the predictive performance of climatological forecasts based on a 350-day rolling training period was also tested. However, such a long training period is unlikely to capture the seasonal variations in visibility. In terms of the mean CRPS and RMSE of the mean, this forecast is far behind the various MLP and POLR models, its mean absolute deviation in 90 % coverage from the required value is 3.90 %, combined with an average width similar to the MLP approaches. The only score where the performance of the 350-day climatology is reasonable is the LogS, at 192 h, 216 h and 240 h it can compete with the regional versions of both studied post-processing techniques.

Appendix C

Dawid-Sebastiani score

As mentioned in Chapter 6, the comparison of multivariate post-processing methods described in Chapter 2, alongside the energy and variogram scores, was also evaluated using the Dawid-Sebastiani score (DSS; Dawid and Sebastiani, 1999), which is based on the first two moments of the prediction. The related results are presented in detail in this appendix.

C.1 Definition

For an L -dimensional forecast ensemble $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_M$ and vector $\mathbf{y}$ the DSS is defined as

$$\text{DSS} := \ln(\det \mathbf{S}) + (\mathbf{y} - \bar{\mathbf{f}})^\top \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{f}}),$$

where $\bar{\mathbf{f}}$ and $\mathbf{S}$ denote the ensemble mean and covariance matrix, respectively.

To handle the possible singularity of the sample covariance matrix, Wilks (2020) suggests to use a regularized version of the DSS (RDSS) defined as

$$\text{RDSS} := \min_{\lambda} \left[\ln(\det \boldsymbol{\Theta}^{-1}(\lambda)) + (\mathbf{y} - \bar{\mathbf{f}})^\top \boldsymbol{\Theta}(\lambda) (\mathbf{y} - \bar{\mathbf{f}}) \right],$$

where $\boldsymbol{\Theta}(\lambda)$ is a sparse approximation of $\mathbf{S}$ corresponding to regularization parameter λ obtained using graphical lasso (Friedman *et al.*, 2008).

C.2 Forecast evaluation with RDSS

Following the suggestions of Wilks (2020), for all three investigated weather quantities we calculated the RDSS by optimizing for λ in the interval $[0.01, 0.02]$, while $\boldsymbol{\Theta}(\lambda)$ was obtained with the help of the `glasso` R package. For simplicity we considered just the raw ensemble and multivariate forecasts addressing

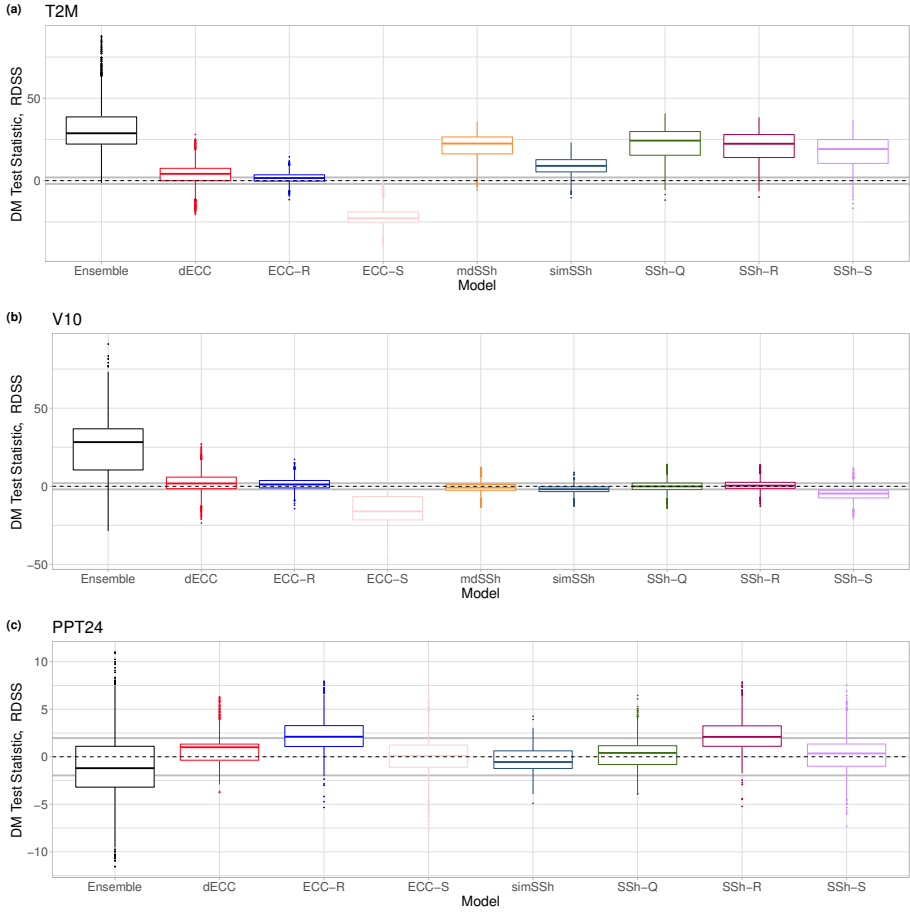


Figure C.1: Boxplots of the DM test statistic investigating the significance of the difference in RDSS from the reference ECC-Q method for T2M (a), V10 (b) and PPT24 (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5 % level of significance.

temporal dependence. However, according to Figure C.1, the RDSS results in a rather similar ranking of the various post-processing methods as the ES (see Figures 6.2c, 6.7c, and 6.11c).