UiO : University of Oslo

Nina Schuhen

Statistical post-processing of weather forecast ensembles: obtaining optimal deterministic and probabilistic predictions at multiple time scales

Thesis submitted for the degree of Philosophiae Doctor

Department of Geosciences

Faculty of Mathematics and Natural Sciences

Norsk Regnesentral

UNIVERSITAS OSLOENSIS MDCCCXI

2020

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2339

ISSN *1501-7710*

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Preface

This thesis is submitted in fulfilment of the requirements for the degree of *Philosophiae Doctor* at the University of Oslo. The research presented here is conducted under the supervision of Dr. Thordis L. Thorarinsdottir and Prof. Dr. Frode Stordal.

Acknowledgements

First and foremost, I would like to thank Thordis L. Thorarinsdottir, without whom this thesis would not have been possible. I am also grateful to Alex Lenkoski for conceiving the idea behind RAFT.
Special thanks go to everyone at Norsk Regnesentral for three wonderful years and to people at CICERO for their support during the last months.

Nina Schuhen
Oslo, November 2020

List of papers

Paper I

Schuhen, N., Thorarinsdottir, T. L. and Lenkoski, A. (2020). 'Rapid adjustment and post-processing of temperature forecast trajectories'. In: *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 727, pp. 963-978. DOI: 10.1002/qj.3718

Paper II

Schuhen, N. (2020). 'Order of operation for multi-stage post-processing of ensemble wind forecast trajectories'. In: *Nonlinear Processes in Geophysics*, vol. 27, no. 1, pp. 35-49. DOI: 10.5194/npg-27-35-2020

Paper III

Thorarinsdottir, T. L. Schuhen, N. and Lenkoski, A. (2020). 'Trajectory adjustment of lagged seasonal forecast ensembles'. *Technical report 19/20. Norsk Regnesentral.*

Paper IV

Thorarinsdottir, T. L. and Schuhen, N. (2018). 'Verification: assessment of calibration and accuracy'. In: *Statistical postprocessing of ensemble forecasts*. Ed. by Vannitsem, S., Wilks, D. S. and Messner, J. W. Elsevier. Chap. 6, pp. 155-186. DOI: 10.1016/b978-0-12-812372-0.00006-6

Contents

Preface i

List of papers iii

Contents v

List of Figures vii

List of Tables xi

1 Introduction 1

1.1 Numerical weather prediction and ensembles 1

1.2 Sub-seasonal to seasonal forecasting 3

1.3 Optimising forecast skill 5

2 Statistical post-processing 9

2.1 Univariate post-processing 9

2.2 Multivariate post-processing 14

3 Rapid adjustment of forecast trajectories 17

3.1 RAFT for ensemble mean forecasts 17

3.2 RAFT for ensemble members 30

3.3 Order of operation for post-processing of multivariate forecasts . . 33

3.4 Forecast jumpiness and consistency 36

3.5 Seasonal temperature forecasts 37

4 Forecast verification 47

4.1 Univariate forecast verification 47

4.2 Verification of forecasts for extremes 52

4.3 Multivariate forecast verification 53

4.4 Comparing probability distributions 56

5 Summary of papers 59

6 Conclusions 65

References 69

Papers 84

I Rapid adjustment and post-processing of temperature forecast trajectories 85

II Order of operation for multi-stage post-processing of ensemble wind forecast trajectories 103

III Trajectory adjustment of lagged seasonal forecast ensembles 121

IV Verification: assessment of calibration and accuracy 137

List of Figures

3.1 Figure 1 from Paper I. Diagram of a typical forecast cycle, where new model runs (FC1, FC2) are initialised every six hours and forecast lead times are one hour apart. The MOGREPS-UK version used here is configured in this way. 17

3.2 (a) Verification rank histogram of the raw MOGREPS-UK surface temperature forecasts from the 15 UTC run and PIT histogram of the corresponding EMOS temperature forecasts. (b) Verification rank histogram of the raw MOGREPS-UK surface wind speed forecasts from the 15 UTC run and PIT histograms of the forecasts post-processed with EMOS, using truncated Gaussian (gEMOS) and truncated logistic distributions (logEMOS). 19

3.3 (a) Figure 4a from Paper I. Empirical correlation coefficient of the temperature mean forecast error for every lead time combination of the 03 UTC model run at Heathrow Airport during the training period. Only correlations significant at the 10% level are shown. (b) Figure 4a from Paper II. Empirical correlation coefficient of the wind speed mean forecast error for every lead time combination of the 15 UTC model run at The Cairnwell during the training period. Only correlations significant at the 10% level are shown. 22

3.4 Figure 5 from Paper I. Diagram of a forecast cycle for an hourly forecast issued every 6 hours, with RAFT applied as new observations become available. Forecasts in grey are only used as predictors by means of their observed error and are not adjusted themselves. 24

3.5 Figure 9 from Paper I. (a) RMSE of the EMOS and RAFT mean temperature forecasts over lead time. The scores are averaged over all dates and locations in the test period for model runs initialised at 21 UTC. RAFT error corrections are carried out only once at lead time $t + 1$. (b) is as (a), but RAFT is carried out for all lead times until the end of the trajectory. 25

3.6 Figure 4 from Paper II. RMSE over lead time for gEMOS (red solid line) and logEMOS (red dashed line) mean wind speed forecasts, as well as their RAFT adjustments (blue solid and blue dashed lines, respectively). The scores are averaged over all dates and locations in the test period for model runs initialised at 15 UTC. (a) RAFT is only carried out until the adjustment at $t + 15$. (b) RAFT is carried out until its last iteration at $t + 35$. 26

3.7 (a) As Figure 3.5b, with added RMSE scores of the most current EMOS mean forecasts for each lead time, i.e., the first 6 hours of the most recent model run. (b) As Figure 3.6b, with added RMSE scores of the most current EMOS mean forecasts for each lead time, i.e., the first 6 hours of the most recent model run. 28

3.8 (a) PIT histograms of EMOS and RAFT mean temperature forecasts, aggregated over all sites and dates from model runs initialised at 15 UTC in the test data set. RAFT forecasts are taken from the final adjustment for each lead time. (b) PIT histograms of gEMOS and RAFT mean wind speed forecasts, aggregated over all sites and dates from model runs initialised at 15 UTC in the test data set. RAFT forecasts are taken from the final adjustment for each lead time. 29

3.9 (a) Correlation matrix of an EMOS ensemble member for the 15 UTC run at The Cairnwell, with only correlations significant at the 10% level shown. (b) Corresponding RAFT$_{\text{ens}}$ adjustment period length for each lead time and the same ensemble member. 31

3.10 Verification rank histogram of the gEMOS and logEMOS + RAFT$_{\text{ens}}$ forecasts, aggregated over all dates, locations and lead times in the evaluation set where the NWP model was initialised at 15 UTC. RAFT forecasts have been adjusted using the observations recorded 2 hours earlier. 32

3.11 Figure 7a from Paper II. Mean CRPS for every step in the RAFT$_{\text{m}}$ and RAFT$_{\text{ens}}$ process. Scores are averaged over all lead times, sites and dates for the 15 UTC model initialisation time. 33

3.12 Figure 6 from Paper II. Average rank histograms for different combinations of post-processing methods. Data points are aggregated over all sites, model runs and lead times. All RAFT forecasts have been adjusted using the observation measured 2 hours earlier. 35

3.13 Band depth histogram for different combinations of post-processing methods. Data points are aggregated over all sites, model runs and lead times. All RAFT forecasts have been adjusted using the observation measured 2 hours earlier. 36

3.14 Convergence index for the EMOS and EMOS + $\mathrm{RAFT_m}$ mean temperature forecasts as a function of the ensemble mean forecast convergence index, averaged for every observation station over all dates in the test data set. RAFT scores are taken from the final adjustment for each lead time. 38

3.15 Correlations between forecast anomaly errors of the ensemble mean at monthly lead times for the model run initialised on 1 May. All correlations are significant at the 10% level. 39

3.16 Figure 2 from Paper III. (a) Correlations between forecast anomaly errors of the ensemble mean at weekly lead times for the model run initialised on 1 May. All correlations are significant at the 10% level. (b) The resulting adjustment periods for each forecast lead time. 40

3.17 Figure 3 from Paper III. RMSE skill scores for five different runs of GloSea5 compared against ERA5 climatology. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015. The original GloSea5 forecasts are indicated with dashed lines, while RAFT forecasts updated one week prior to the realisation time are indicated by solid lines. 41

3.18 Figure 4 from Paper III. RMSE skill scores for a comparison against ERA5 climatology for the combination of all five forecast runs (blue dashed line), for RAFT-processed ERA5 climatology of adjustment lead time one week (brown solid line), and for a combination of the RAFT-processed ensemble means for all five forecast runs at an adjustment lead time of one week (blue solid line). For comparison, these skill scores are overlaid on the results shown in Figure 3.17. 42

3.19 Anomaly correlation coefficient of different mean forecasts and the ERA5 reanalysis. Shown are the GloSea5 individual runs and their RAFT adjustments in the background and the combined GloSea5 mean forecast (dashed dark blue line), as well as the combined RAFT mean forecast (solid dark blue line). All RAFT forecasts have one week adjustment lead time. 43

3.20 Figure 5 from Paper III. RMSE skill scores for various model combinations of forecasts issued in week five compared against ERA5 climatology. Each combination consists of the most recently available runs. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015. 44

4.1 Figure 6.7 from Paper IV. Top row: Mean absolute error, CRPS and ignorance score, and the 95% bootstrap confidence interval for the five forecast distributions, if the true distribution is a Gumbel distribution. Scores are based on 1000 forecast-observation pairs. Bottom row: Same as above, but scores are based on 1 million forecast-observation pairs. . 51

4.2 Figure 6.11 from Paper IV. Mean scores and 95% bootstrap confidence interval for the four versions of the CRPS. Top row: twCRPS with weight functions w_2 and w_3. Bottom row: CRPS restricted to observations above the threshold u and unweighted CRPS. 54

List of Tables

3.1 Continuous ranked probability score and root-mean-square error (in °C) of raw and post-processed MOGREPS-UK temperature forecasts, averaged over all forecast cases from the 15 UTC run in the test set. All pairwise score differences are statistically significant at $\alpha = 0.01$. 20

3.2 Continuous ranked probability score and root-mean-square error (in knots) of raw and post-processed MOGREPS-UK wind speed forecasts, averaged over all forecast cases from the 15 UTC run in the test set. All pairwise score differences are statistically significant at $\alpha = 0.01$. 20

3.3 Continuous ranked probability score and root-mean-square error (in knots) of different combinations of EMOS and RAFT methods, averaged over all forecast cases from the 15 UTC run in the test set. The RAFT forecasts are taken from the final adjustments for each lead time. All pairwise score differences are statistically significant at $\alpha = 0.05$, apart from the RMSE differences between identical combinations using gEMOS and logEMOS. 32

3.4 Multivariate scores for different combinations of post-processing methods, averaged over all locations, lead times and dates in the test set, where the model was initialised at 15 UTC. RAFT forecasts are taken from the final adjustment for each lead time. The variogram score was calculated with order 0.5, as per the recommendation in Scheuerer and Hamill (2015). All score differences are significant at the 5% level. 34

Chapter 1

Introduction

When making decisions based on predictions about the future, it is imperative to consider information about the prediction's uncertainty. This is particularly true for weather forecasting, where the chaotic nature of the atmosphere does not allow for perfect precision. It has been shown that having access to uncertainty information leads to better decisions, especially in the case of extreme or severe events (Joslyn and LeClerc, 2012). Nonetheless, the estimated uncertainty needs to be accurate and informative. Post-processing weather forecasts with statistical methods improves conventional weather models by reducing biases and calibrating predictive probability distributions. In this thesis, we will introduce rapid adjustment of forecast trajectories (RAFT), a new post-processing approach that utilises the error correlation between lead times to update parts of a forecast trajectory that have not yet realised.

1.1 Numerical weather prediction and ensembles

Numerical weather prediction (NWP) originated with Vilhelm Bjerknes, who in 1904 suggested that weather forecasts can be obtained by applying the governing equations of fluid dynamics and integrating the current state of the atmosphere forward in time (Bjerknes, 1904). In 1922, Lewis Fry Richardson made the first attempt at producing such a forecast (Richardson, 1922), although it took him at least several weeks and the result was far from accurate (Lynch, 2006). His dream was to one day have enough (human) computers available to keep up with the current development of weather. The first successful numerical weather forecast came with the arrival of ENIAC, the first electronic multi-purpose digital computer, and the experiments conducted by Charney, Fjørtoft, and von Neumann (1950). They managed to accurately forecast the geopotential height over North America 24 hours ahead, even if the computations took longer than a day to finish (Lynch, 2008). Due to the rapid increase in computing power, the first global spectral model became fully operational in the 1980s (e.g., Kalnay, 2002).

A major paradigm shift in NWP occurred following Edward Lorenz's first ventures into chaotic systems (Lorenz, 1963; Lorenz, 1969), where he established that the atmosphere's development is critically sensitive to its initial state, of which a perfect representation can never be achieved. This limits the predictability of weather phenomena to a maximum of about two weeks, as small-scale errors grow so rapidly that they carry over to larger scales after a few days. Even with unlimited computing power and perfect numerical models, it is still impossible to produce perfect forecasts of a deterministic chaotic system.

Although Edward Epstein (Epstein, 1969) proposed a stochastic-dynamic solution by treating the weather variables as random and generating multiple predictions using the Monte Carlo method (Metropolis and Ulam, 1949), the required computational resources were far from realistic (Lewis, 2005). Leith (1974) found that the predictive mean can be accurately determined with 8 samples, but higher moments and thus much larger sample sizes are needed for modelling more complex distributions (Lewis, 2014).

It took another two decades until operational forecast ensembles as we know them today were launched by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) in 1992 (e.g., Molteni et al., 1996; Tracton and Kalnay, 1993). An ensemble prediction system comprises of multiple runs of a numerical forecasting model, each differing slightly in the specifications of the initial conditions or the model parameters, or both. In order to make the most use of the ensembles, it is important to choose sets of initial conditions that, when integrated forward, span a large part of all possible atmospheric states. At the ECMWF, singular vectors, pointing towards the fastest growth over a finite time interval, were initially used for this purpose (Buizza and Palmer, 1995; Buizza et al., 1993), while the NCEP model relied on bred vectors, where scaled perturbations are frequently added to the control run of the non-linear model (Toth and Kalnay, 1993).

Ensemble forecasting systems consist of several stages. First, the current state of the atmosphere, the analysis, has to be determined from recent observations. This process, called data assimilation, is also used to create reanalyses, i.e., historical atmospheric states that are employed for climate monitoring and verification. One such reanalysis data set is ERA5 (Copernicus Climate Change Service, 2017), produced at ECMWF based on their model IFS.

While some models combine a single best analysis with the ensemble perturbations to form the initial conditions, others incorporate ensemble techniques into the data assimilation process (for an overview see e.g., Hamill, 2006) and thereby try to obtain more accurate information about the expected forecast error. The German Weather

Service (DWD), for example, uses a local ensemble transform Kalman filter approach for their COSMO (COnsortium for Small-scale MOdelling) convective-scale model (Schraff et al., 2016).

Another fundamental way to create useful ensemble perturbations is to incorporate uncertainty in the second stage of the forecasting process, when the analysis is integrated forward in time to create predictions. The model itself is imperfect, with some physical processes not resolved directly but through parameterisations. These, in combination with spatial and temporal discretisation due to mathematical limitations, constitute further sources of uncertainty. Popular methods developed to account for the resulting errors include stochastically perturbing the parameterisation tendencies (e.g., Buizza, Miller, and Palmer, 1999) or modelling missing physics processes with a stochastic kinetic energy backscatter scheme (e.g., Tennant et al., 2011).

Many forecast providers also run limited-area weather forecasting models, which have a higher grid resolution over a certain region. These models are often convection-permitting and thus able to resolve processes on a much finer scale than the global models. In Paper I and Paper II, data from the UK Met Office's MOGREPS-UK (Met Office Global and Regional Ensemble Prediction System) limited-area ensemble are used, which has been in operation since 2012 (Hagelin et al., 2017). For such models, the initial and boundary conditions are usually taken from one or more global models and sometimes combined with perturbations of the model physics. In the case of the MOGREPS-UK model, the global ensemble MOGREPS-G originally provided the initial and boundary conditions, however since a major upgrade in 2016, the initial conditions have been created by using a combination of the MOGREPS-G perturbations and the analysis of the high-resolution deterministic UK variable-resolution (UKV) model. Recently, MOGREPS-UK received another significant change. Instead of producing twelve ensemble member forecasts four times a day for the next 36 hours (as in the data set here used), the ensemble is now run on an hourly cycle with three ensemble members and predictions for the next 120 hours. The forecasts from the six most recent model cycles are collected and comprise an 18-member lagged ensemble (Met Office, 2020).

1.2 Sub-seasonal to seasonal forecasting

Although the predictability limit for weather lies at about two weeks (Zhang et al., 2019), some long-range phenomena can still provide useful indications about the average weather over longer time periods, from a few weeks to a whole season. Sub-seasonal to seasonal forecasting tries to bridge the gap between weather and climate predictions by looking at how these phenomena possibly influence deviations from climatology for certain regions and weather parameters. For example, sea surface temperature changes in the Pacific Ocean, the El Niño-Southern Oscillation (ENSO), are related to abnormal rainfall activity across the globe.

One of the biggest drivers of seasonal variability in Europe is the North Atlantic Oscillation (NAO), a teleconnection pattern defined as the difference in sea level pressure between two points in the North Atlantic (Feldstein and Franzke, 2017). A strongly positive NAO index often coincides with warm and wet winters in northern and cold and dry winters in southern Europe. However, recent results suggest that predictability for NAO, and thus seasonal anomalies, shows some multi-decadal variability (Weisheimer et al., 2017), making useful predictions potentially difficult during future periods of low predictability.

In order to incorporate long-range processes, atmospheric general circulation models are coupled with ocean, land-surface and sea-ice models (Vitart and Robertson, 2019). One such system is the UK Met Office's GloSea5 (Global Seasonal forecast system version 5), which is based on the HadGEM3 family of climate prediction models (MacLachlan et al., 2015). The underlying atmospheric model of both GloSea5 and the short-range MOGREPS-UK is the Unified Model, designed to produce seamless forecasts from hours to months (Met Office, 2019).

Ensembles play a particularly important role in seasonal weather forecasting, as forecast uncertainty increases with lead time and seasonal forecasts are often given as likelihoods of deviation from the average weather. To get the most benefit from the computational resources available, GloSea5 uses a lagged ensemble approach: every day, four ensemble members are initialised from the most recent analysis, two of them running out to 210 days. The forecasts from the last three weeks are then combined to form a 42-member seasonal ensemble (MacLachlan et al., 2015). In addition to the operational forecasts, produced since 2013, a set of hindcasts with three ensemble members is generated every week, covering the same week over a 14-year-period in the past. The purpose of this data set is to provide training data for bias-correcting the operational forecasts, with the hindcast data characteristics matching those of the current model version.

While many sources of forecast skill remain theoretical and relationships between teleconnection patterns still need to be fully understood (Hoskins, 2012; Robertson and Vitart, 2019), seasonal prediction models are rapidly evolving and thus rising in value for many customers (White et al., 2017). Applications include the energy sector, particularly renewable energies (e.g., Orlov, Sillmann, and Vigo, 2020), agriculture (e.g., Klemm and McPherson, 2017) and public health (e.g., MacLeod et al., 2015).

1.3 Optimising forecast skill

In order to be considered skilful, forecasts need to be accurate. For deterministic forecasts, where usually only one single number is predicted, this means that this number should be as close to the observation as possible. In the probabilistic case, the forecast can take a number of different forms, like a percentile, an exceedance probability, a confidence interval or a full probability distribution.

There are two important concepts that in combination describe the skill of a probabilistic forecast: calibration is the statistical consistency between the forecasts and the observations, whereas sharpness refers to the concentration of the forecast uncertainty (Gneiting, Balabdaoui, and Raftery, 2007). A forecast is calibrated if an event that is predicted with a certain probability p on average transpires in p percent of all forecast cases. NWP ensemble forecasts are typically underdispersed, meaning that the ensemble spread is too small and the model is too confident (e.g., Hamill and Colucci, 1997). Forecasts derived from climatology on the other hand are by design well-calibrated, but not sharp and therefore potentially not very useful.

Optimal sharpness constitutes a 100% confident probabilistic forecast, e.g., when a confidence interval has width zero. This is, however, only desired if this confidence is justified, i.e., if the forecasts are reliable. Gneiting, Balabdaoui, and Raftery (2007) therefore introduce the paradigm of maximising the sharpness of the predictive distribution subject to calibration. Given a choice of reasonably calibrated forecasts, one should always choose the one producing the sharpest predictions. A collection of tools assessing both calibration and sharpness are described in Chapter 4 and in Paper IV. The latter also provides studies investigating the behaviour of these tools (e.g., for limited sample sizes), as well as suggestions for best practice.

Despite all of the efforts in designing and improving NWP models as described in Section 1.1, ensembles are not able to capture all of the uncertainties involved in weather forecasting (Raftery et al., 2005). They are e.g., limited in their spatial and temporal resolution, number of ensemble members, representation of atmospheric processes and accuracy of the initial state.

For this reason, a multitude of statistical post-processing methods have been developed over the past decades, aiming to correct the deterministic and probabilistic forecasts generated by NWP models and thus account for some of the missing atmospheric uncertainty. During this process, the models are usually compared against observations, either on a grid or at specific sites. In the latter case, it is also possible to account for local effects that might not be present in the model, e.g., due to a large difference between the model orography and the actual terrain.

Although NWP models are continuously improved and benefit strongly from the rapid increase in available computing power, Hemri et al. (2014) show that the benefit gained from statistical post-processing remains almost constant, even if the underlying model's skill increases. Consequently, there is and will be a need for statistical post-processing in order to create reliable and useful forecast products from the raw NWP output.

While it is essential to make forecasts as accurate and as reliable as possible, they also have to be of relevance to customers, who may include the public and commercial sectors, as well as the general public. This corresponds to the type 3 criterion for the goodness of a weather forecast described in Murphy (1993). Human forecasters add a large amount of value to numerical weather forecasts and are essential for giving warnings for high-impact weather events (e.g., Novak et al., 2011). However, it is impossible to have forecasters manually assess and interpret every time series or map and websites of most forecast providers are instead supplied with direct output from the numerical weather forecasting and post-processing systems. This makes the forecasts somewhat susceptible to inconsistencies in time and space if locations and forecast lead times are post-processed separately. Some of these issues, including inconsistencies between weather parameters, can partially be addressed with multivariate post-processing as described in Section 2.2.

Furthermore, forecasts are usually updated only when a new NWP model run has finished, which can range from hourly to twice-daily updates. In the meantime, it might become obvious that the current forecast run suffers from a systematic error as soon as the first few observations are recorded. For instance, the cloud cover could be overestimated for the next six hours and the temperature therefore underestimated. The customer then sees a cloudy forecast that fails to materialise in reality and is possibly replaced with a more accurate forecast some hours later. This discrepancy between the current forecast and the short-term outcome may result in a loss of confidence in the forecast provider, in addition to any loss due to decisions made on the basis of an inaccurate forecast.

With the new RAFT post-processing framework, we want to provide a solution to this issue. Forecast trajectories can now receive frequent updates every time a new observation becomes available, based on the correlation between the observed errors at consecutive lead times. This means that systematic errors can be caught early and corrected for several hours ahead when applied to short-range ensembles like MOGREPS-UK. Consequently, forecasts from older model runs with large lead times become more valuable and can outperform the first few forecasts from the newest run. This is especially important as models take multiple hours to run and are not available until some time after initialisation. In Paper I and Paper II, we illustrate how RAFT is applied to short-range temperature and wind speed forecasts, both deterministic and probabilistic.

RAFT can also provide a large advantage to ensembles with a rapid update cycle, where forecasts from several runs are combined to a large ensemble. As these forecasts correspond to different lead times, their relative skill can vary substantially. With RAFT, all ensemble members are updated using the most recent error information and the differences in skill are balanced out. The application of RAFT in such a setting is discussed in Paper III, although there are limitations to its capabilities due to the lack of correlation between lead times of seasonal forecasts.

Chapter 2

Statistical post-processing

In the following, we describe how statistical post-processing can be used to improve the skill of deterministic and probabilistic forecasts. We differentiate between univariate post-processing, where usually only one location, lead time and variable is addressed, and multivariate post-processing, which incorporates dependencies between multiple dimensions. The methods in this chapter are applied in Paper I and Paper II to provide a baseline for RAFT.

2.1 Univariate post-processing

As a relatively sparse way to correct for deterministic and probabilistic biases, there are now several post-processing methods in operational use around the world, designed for different weather variables and forecasting scenarios. Hess (2020) illustrate how such an operational framework can work: Deutscher Wetterdienst employ a combination of the well-known model output statistics approach (MOS; Glahn and Lowry, 1972) for deterministic forecasts of individual variables and event probabilities, as well as logistic regression (Hosmer, Lemeshow, and Sturdivant, 2013) for more complex probabilities.

There are two well-established statistical post-processing methods for ensemble forecasts that produce full probability distributions, from which any deterministic or probabilistic forecast can be derived: Bayesian Model Averaging (BMA; Raftery et al., 2005) and Ensemble Model Output Statistics (EMOS; Gneiting et al., 2005). BMA dresses each ensemble member forecast with an appropriate probability distribution, e.g., Gaussian distributions in the case of temperature forecasts, and then combines these distributions to a weighted average mixture distribution. The ensemble member forecasts are first bias-corrected using linear regression and the individual weights and variance of the mixture distribution then estimated with maximum likelihood (Fisher, 1922) and the expectation-maximisation algorithm (Dempster, Laird, and Rubin, 1977; McLachlan and Krishnan, 2008). The weights for the individual ensemble member distributions can be interpreted as the relative forecast skill of that member over the training period. Further variants of BMA have been developed for wind speed (Baran, 2014; Sloughter, Gneiting, and Raftery, 2010), precipitation (Schmeits and Kok, 2010;

Sloughter et al., 2007) and wind vectors (Sloughter, Gneiting, and Raftery, 2013), as well as a method for jointly post-processing temperature and wind speed (Baran and Möller, 2014).

While BMA has the ability of forming multimodal predictive distributions, which is especially advantageous for ensembles whose members can be grouped into clusters or weather scenarios, it is relatively expensive to run – although the computational cost of any statistical post-processing is negligible compared to running the NWP model itself. The much more sparse EMOS method only fits a single unimodal predictive distribution, but is conceptually simpler and therefore easier to adapt and faster to compute. In terms of forecast skill, both methods usually perform on a similar level (e.g., Baran, Horányi, and Nemoda, 2013) and we thus prefer to use EMOS (sometimes also called non-homogeneous regression) in Paper I and Paper II.

Like many post-processing methods, EMOS is based on the idea that the ensemble will provide a flow-dependent estimate of the uncertainty in a given weather scenario, it is just on average too confident and needs to be calibrated. Therefore a single standard probability distribution is fitted across all ensemble members and the distribution characteristics are modelled as functions of the ensemble. Again, the distribution depends on the type of weather variable at hand. A variety of EMOS versions are available for temperature (Gneiting et al., 2005; Scheuerer and Büermann, 2014), wind speed (e.g., Baran and Lerch, 2016; Scheuerer and Möller, 2015; Thorarinsdottir and Gneiting, 2010), wind gust (Thorarinsdottir and Johnson, 2012), precipitation (e.g., Baran and Nemoda, 2016; Scheuerer, 2013; Scheuerer and Hamill, 2015), wind vectors (Schuhen, Thorarinsdottir, and Gneiting, 2012) and combined wind speed and temperature (Baran and Möller, 2016).

In Paper I, temperature forecasts are post-processed in a conventional manner using Gaussian distributions for modelling the EMOS predictive distribution, as in Gneiting et al. (2005), before applying the new RAFT method. For every location, forecast run and lead time, there are ensemble forecasts $X_1, \ldots, X_m$, which correspond to an observation Y. We model the predictive distribution of the observation, conditional on the ensemble members, as

$$Y \mid X_1, \ldots, X_m \sim \mathcal{N}\left(\mu, \sigma^2\right), \tag{2.1}$$

where $\mathcal{N}$ is a Gaussian distribution with mean μ and variance σ^2. The mean and variance are connected to the ensemble mean $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ and the ensemble variance $S^2 = \frac{1}{m}\sum_{i=1}^{m}\left(X_i - \bar{X}\right)^2$ in the following manner:

$$\mu = a + b^2 \cdot \bar{X} \tag{2.2}$$

$$\sigma^2 = c^2 + d^2 \cdot S^2. \tag{2.3}$$

Here, we treat the ensemble members as exchangeable and only use the ensemble mean $\bar{X}$ as predictor for the EMOS mean μ. It is also possible to include the individual ensemble members in the regression equation (2.2), as well as other potentially useful predictors. However, this increases the amount of necessary training data and can lead to overfitting. The coefficient b is squared to make it easier to interpret, while c and d in (2.3) are squared in order to ensure a positive variance. The latter also provide an indicator of the performance of the ensemble over the training period. If c is large and d close to zero, the ensemble spread is not a good forecast of the actual forecast uncertainty and is disregarded; in the case of c being small and d close to 1, the ensemble spread reflects the actual predictability over the training period.

There has been some discussion as to the optimal estimation of the coefficients a, b, c and d. Gneiting et al. (2005) originally propose to minimise the continuous ranked probability score (CRPS; see Section 4.1), a proper scoring rule that optimises calibration and sharpness at the same time. Another option is maximum likelihood estimation, which can also be interpreted as minimising a proper scoring rule, namely the ignorance score. Gebetsberger et al. (2018) show that although both methods perform similarly overall, CRPS minimisation works slightly better if the forecast distribution is not perfectly specified and the forecasts are ultimately verified using the CRPS. Therefore we use this method in Paper I and Paper II to obtain the estimated parameters for the EMOS predictive distributions.

The parameters are calculated based on a training set, i.e., forecast-observation pairs over a given time period. While some statistical applications allow for a fixed, longer training period, post-processing of weather forecasts often involves a rolling training period consisting of data from the last n days preceding the model run of interest (e.g., Gneiting et al., 2005). In order to correct the specific deficiencies of NWP ensembles, the training data and the current forecasts should ideally have been generated using the same model setup. With operational NWP models, frequent changes are common and therefore a long time period of past data from every model generation (reforecasts; Hamill, 2018) pose an enormous computational burden. In cases where such reforecasts are not feasible, as with the MOGREPS-UK model in Paper I and Paper II, it is better to use a relatively short rolling training period so that the transition between model generations is smooth. It has the additional advantage that errors associated with particular weather regimes can be captured in a flow-dependent manner. However, due to the smaller amount of data, coefficients can exhibit a jumpy behaviour, which in turn might lead to a loss in predictive skill. Lang et al. (2020) have recently shown that sliding training period windows that incorporate at least some historical data have an advantage over the conventional approach. In an interesting proof-of-concept, Demaeyer and Vannitsem (2020) investigate a technique based on response theory to incorporate small model changes into the estimation of post-processing coefficients.

When selecting an appropriate set of training data, it is necessary to choose between two approaches, which Thorarinsdottir and Gneiting (2010) call regional and local EMOS. The former pools the training data from all locations or grid points and then estimates only one set of coefficients. This results in more stable estimates and therefore shorter training period lengths, in addition to the possibility of applying these coefficients to locations without available observation data. The local approach uses only training data from the location or grid point at hand, which requires longer training periods, but often produces more skilful forecasts that take into account local effects.

After estimating the parameters by minimising the CRPS over the training data, we then plug the ensemble mean and variance from the current forecast run in (2.2) and (2.3), respectively. From the resulting predictive distribution, we can generate any deterministic or probabilistic forecast product. This process is repeated for every forecast run, lead time and, in the case of local EMOS, location. In order to avoid large jumps in the coefficients from day to day, the previous day's final estimates are used as starting values for the optimisation algorithm. Occasionally, in particular when d is close to zero, this can lead to the algorithm being stuck in a local minimum. In such cases, we reset the starting values for c and d.

While the Gaussian distribution is usually a very good fit for temperature forecast, it is more difficult to find a suitable distribution for wind speed, as the values are non-negative and the tails can be quite heavy, depending on the location. For instantaneous wind speed, Thorarinsdottir and Gneiting (2010) propose a truncated Gaussian distribution that is cut off at zero. This means that any negative value has probability zero and the remaining part of the distribution is adjusted accordingly. Other distributions commonly used are truncated logistic distributions, gamma distributions (both Scheuerer and Möller, 2015), log-normal distributions (Baran and Lerch, 2015) and generalised extreme value distributions (Lerch and Thorarinsdottir, 2013). In Paper II, we will apply the truncated Gaussian (gEMOS) and truncated logistic (logEMOS) models to generate post-processed forecasts as baseline for the new RAFT method. The two wind speed EMOS models are defined as

$$Y \mid X_1, \ldots, X_m \sim \mathcal{N}^+ \left(\mu, \sigma^2\right) \tag{2.4}$$

for gEMOS, with μ and σ^2 being the location and scale parameters of the truncated Gaussian distribution $\mathcal{N}^+$, and

$$Y \mid X_1, \ldots, X_m \sim \mathcal{L}^+ \left(\mu, s\right) \tag{2.5}$$

for logEMOS, where μ is the location parameter and s the scale of the truncated logistic distribution $\mathcal{L}^+$.

Again, we want to link the distribution characteristics to the ensemble mean $\bar{X}$ and variance S^2:

$$\mu = a + b^2 \cdot \bar{X} \tag{2.6}$$

$$\sigma^2 = c^2 + d^2 \cdot S^2. \tag{2.7}$$

Note that here we model the location parameter μ and not the mean, and in the case of gEMOS the scale parameter σ^2 and not the variance. For logEMOS, (2.7) refers to the variance $\sigma^2 = s^2\pi^2/3$. As the CRPS is available in a closed form for both truncated distributions (see Section 4.1), parameter estimation is straightforward and can be conducted in the same manner as for temperature.

Although BMA and EMOS are the most popular, there is a wide variety of post-processing methods, catering to specific weather variables and forecasting needs. For example, some applications might require a non-parametric approach, where the individual ensemble members are corrected instead of constructing a predictive distribution. Comprehensive overviews can be found in Wilks (2018) and Vannitsem et al. (2020). The latter article also summarises current operational applications of statistical post-processing – including forecast blending – and their challenges, as well as potential future research directions. A prominent new area of research mentioned is the application of machine learning in post-processing. Developed by NCAR (National Center for Atmospheric Research), the DICAST system, combining statistical post-processing and blending of numerous data sources with machine learning, has already been operational for more than two decades and is used for many applications in various sectors, such as wind and solar energy or agriculture (Haupt et al., 2018).

2.2 Multivariate post-processing

The post-processing methods described in the previous section are designed in such a manner that they can be applied to single forecasts valid at a single point in time at a single location. Due to the purely statistical nature of these methods, physical relationships are rarely taken into account and forecasts are potentially no longer consistent. For example, temperatures in summer are usually correlated with cloud cover, however separate post-processing of the two variables can lead to them being completely independent. While there may be scenarios where physical consistency is of no importance, typical applications include at least multiple locations or lead times. For the wind energy industry, it is important to have accurate forecasts about the location and timing of frontal systems (Steiner et al., 2017), whereas in a hydrological context, run-off scenarios require coherent spatio-temporal structures (Hemri, Lisniak, and Klein, 2015).

Some multivariate post-processing methods are tailored to specific applications, such as producing calibrated bivariate forecasts for wind vectors (Pinson, 2012; Schuhen, Thorarinsdottir, and Gneiting, 2012; Sloughter, Gneiting, and Raftery, 2013), either in the form of full predictive distributions or adjusted ensemble forecasts. Hemri, Lisniak, and Klein (2015), e.g., model the correlation structure between different lead times, each first post-processed with EMOS, by using Gaussian copulas. In order to obtain consistent and calibrated forecast fields, Berrocal, Raftery, and Gneiting (2007), Berrocal, Raftery, and Gneiting (2008), and Feldmann, Scheuerer, and Thorarinsdottir (2015) combine BMA and EMOS with the geostatistical output perturbation method (Gel, Raftery, and Gneiting, 2004), also based on Gaussian copulas. These parametric approaches, however, can be quite complicated and expensive to run.

Non-parametric methods that rely on empirical copulas to model the multivariate relationships have proven to be rather versatile and effective when it comes to handling many dimensions or combinations of multiple variables, lead times and locations. There are two state-of-the-art approaches, each with their strengths and weaknesses, depending on the forecasting scenario: ensemble copula coupling (ECC) and Schaake shuffle. They share the same framework, where a specific multivariate dependency template is applied to samples from individually calibrated marginal distributions. The only difference lies in the origin of the dependency template at hand. While ECC (Schefzik, Thorarinsdottir, and Gneiting, 2013) assumes that the raw ensemble correctly portrays the physical relationship between the dimensions of interest, Schaake shuffle (Clark et al., 2004) relies on historical observations.

Due to their design, it is straightforward to combine univariate post-processing with ECC or Schaake shuffle. These methods are computationally efficient, can be applied to forecasts of any dimension and default to the original marginal distributions when evaluated in a univariate manner. In order to apply ECC or Schaake shuffle, we follow these three steps:

1. Apply a univariate post-processing method of choice to the raw ensemble forecasts $X_1^{(d)}, \dots, X_m^{(d)}$ for each dimension d, then draw n samples from the d marginal distributions. Schefzik, Thorarinsdottir, and Gneiting (2013) recommend using equidistant quantiles, as they are optimal with regard to the CRPS (Bröcker, 2012). It is also possible to use other techniques like stratified (Hu et al., 2016) or random sampling. For ECC, the number of samples n is limited to the size of the original NWP ensemble m, while for Schaake shuffle it depends on the number of historic observations selected. From this step, we obtain a set of samples $\widetilde{X}_1^{(d)}, \dots, \widetilde{X}_n^{(d)}$.

2. Extract the dependency template from the selected source. In case of ECC, this is the order statistic of the raw ensemble members, where we note the rank of each ensemble member $X_i^{(d)}$ among the other members $X_1^{(d)}, \dots, X_m^{(d)}$. Any ties are resolved at random. For the Schaake shuffle, we do the same with a set of historical observations $Y_1^{(d)}, \dots, Y_n^{(d)}$. The dependency template is then a permutation $\tau_d(\cdot)$ of the numbers $1, \dots, m$ or $1, \dots, n$ with

$$X_{\tau_d(1)}^{(d)} \leq X_{\tau_d(2)}^{(d)} \leq \dots \leq X_{\tau_d(m)}^{(d)} \quad \text{or} \tag{2.8}$$

$$Y_{\tau_d(1)}^{(d)} \leq Y_{\tau_d(2)}^{(d)} \leq \dots \leq Y_{\tau_d(n)}^{(d)}. \tag{2.9}$$

3. Reintroduce the correlation structure by ordering the post-processed samples $\widetilde{X}_1^{(d)}, \dots, \widetilde{X}_n^{(d)}$ according to the permutation from the previous step. The result is a d-dimensional ensemble with n members that has calibrated marginals and the same relationship between components as the dependency template:

$$\left[\widetilde{X}_{\tau_d(1)}^{(1)}, \dots, \widetilde{X}_{\tau_d(1)}^{(d)}\right], \dots, \left[\widetilde{X}_{\tau_d(n)}^{(1)}, \dots, \widetilde{X}_{\tau_d(n)}^{(d)}\right]. \tag{2.10}$$

ECC has the advantage that no additional data are needed, whereas the Schaake shuffle requires a comprehensive amount of historical observations. While Clark et al.

(2004) initially only use dates from previous years and a small period around the current date, there have been some efforts to develop more sophisticated selection processes, like restricting the data according to similarity conditions (Schefzik, 2016) or matching the marginal distributions of the observations and forecasts (Scheuerer et al., 2017). To circumvent the limitation in sample size for ECC, it is possible to repeatedly sample randomly from the post-processed distributions and apply the reordering multiple times. The aggregated ensemble can then outperform the much smaller ensemble consisting of equidistant quantiles (Wilks, 2014). Ben Bouallègue et al. (2016) combine ECC with the autocorrelation of the forecast error over consecutive lead times in order to improve the representation of temporal dependencies.

In Paper II, we show how the new RAFT method can be combined with ECC to create an optimal post-processing chain. A variant of the EMOS/ECC combination used here, only with a slightly different sampling scheme, can be interpreted as a direct mapping between unprocessed and post-processed ensemble members, often called member-by-member post-processing (Schefzik, 2017). The RAFT$_{\text{ens}}$ method described in Section 3.2 also resembles such member-by-member approaches. Lerch et al. (2020) investigate the relative performance of ECC, Schaake shuffle and a parametric Gaussian copula method (Möller, Lenkoski, and Thorarinsdottir, 2013; Pinson and Girard, 2012) for simulated data, where the forecasts exhibit a variety of misspecifications. A detailed description of multivariate post-processing methods can be found in Schefzik and Möller (2018).

Chapter 3

Rapid adjustment of forecast trajectories

The newly proposed RAFT approach describes a class of statistical post-processing methods, designed to complement the established methods described in the previous chapter. Its goal is to update forecasts using observations that have become available since the model run's initialisation and thus improving the skill of older forecasts, as shown in Paper I and Paper II. For lagged ensembles like in Paper III, RAFT can be used to balance the difference in relative skill between ensemble members. In principle, RAFT applies to forecast scenarios of any range, from short-range to seasonal. However, there are limitations, which we discuss in Section 3.5.

3.1 RAFT for ensemble mean forecasts

The idea behind RAFT is to make use of observations recorded between two model initialisation times to incorporate new information and consequently make the old forecasts more accurate. This prevents forecast products, especially in the short range, becoming outdated and sometimes obviously wrong until the next model run has finished. At this point it is likely that the level of forecast skill suddenly jumps, as the new NWP run replaces the old.

A typical setup of a NWP model cycle can be seen in Figure 3.1, where the model is initialised from a new analysis every 6 hours, producing hourly forecasts for the next few hours to days. We refer to the model runs by their *initialisation time* (in

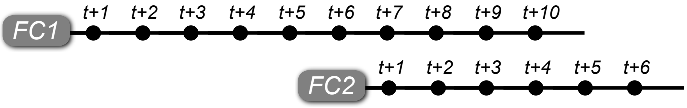


Figure 3.1: Figure 1 from Paper I. Diagram of a typical forecast cycle, where new model runs (FC1, FC2) are initialised every six hours and forecast lead times are one hour apart. The MOGREPS-UK version used here is configured in this way.

UTC) and the time points for which forecasts are produced as *lead times* if they are given relative to the initialisation time. For example, a forecast produced from a run initialised at 03 UTC with lead time 6 (or $t+6$) would here be valid for 09 UTC on the same day. This latter, fixed time point is usually called the *valid* or *validity time.*

The scheme in Figure 3.1 corresponds to the Met Office's MOGREPS-UK ensemble (Hagelin et al., 2017), which is used in Paper I and Paper II. Forecasts are interpolated from the model grid to individual observation locations in the UK and the Republic of Ireland, and corrected for differences between the real and the model orography. In a subsequent process, these data are blended with forecasts from other models such as the global MOGREPS-G and the ECMWF medium-range model, before being published on the Met Office's website (Sharpe, Bysouth, and Stretton, 2017). However, we do not apply RAFT to these blended forecasts, but rather to the individual models.

Our initial goal is to improve the deterministic forecast skill of the ensemble mean forecast to showcase the RAFT framework and then develop a version that also addresses probabilistic forecasts and can be integrated into a comprehensive operational post-processing chain. RAFT does not replace other post-processing methods, but can and should be applied to forecasts that have been subject to conventional post-processing. In this way, methods like EMOS (and ECC, as described in the next section) work in concert with RAFT, but operate at different time scales.

Therefore, we first apply EMOS to temperature and wind speed forecasts from MOGREPS-UK in order to create a baseline for the forecast skill we can obtain from conventional post-processing. Any additional forecast skill is solely achieved by adding information that was not available at the point when the original forecasts were issued. The data set covers a period of about 2.5 years from January 2014 to June 2016. MOGREPS-UK has experienced several operational changes during this time, such as the addition of the high-resolution analysis of the UKV model; however, we can not take these into account here. We use the complete year of 2014 as an estimation period for RAFT and the remaining 1.5 years for evaluation.

All four forecast initialisation times at 03, 09, 15 and 21 UTC are considered, as are all 36 hourly lead times. The data set contains 150 locations for surface temperature and 152 for surface wind speed forecasts, which correspond to weather stations recording SYNOP observations. We can roughly categorise these weather stations as coastal, mountainous and inland sites.

When applying EMOS to the whole data set, as described in Section 2.1, each location, forecast run and lead time are treated separately. We use a rolling training

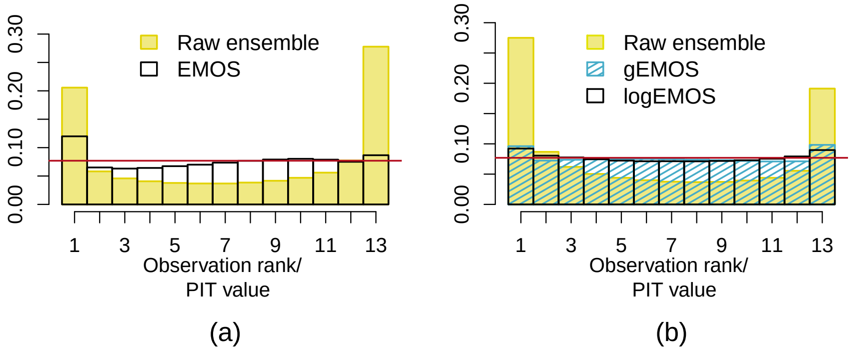


Figure 3.2: (a) Verification rank histogram of the raw MOGREPS-UK surface temperature forecasts from the 15 UTC run and PIT histogram of the corresponding EMOS temperature forecasts. (b) Verification rank histogram of the raw MOGREPS-UK surface wind speed forecasts from the 15 UTC run and PIT histograms of the forecasts post-processed with EMOS, using truncated Gaussian (gEMOS) and truncated logistic distributions (logEMOS).[1]

period of 40 days, which has shown good results in a previous study (Schuhen et al., 2016). This process results in fairly calibrated temperature and wind speed forecasts, as can been seen in Figure 3.2. Here, the verification rank and PIT histograms of the raw ensemble are compared to the post-processed EMOS forecasts for all data in the test set where the model run was initialised at 15 UTC. The MOGREPS-UK forecasts are very underdispersed, in both the temperature and the wind speed case. After applying the EMOS version using Gaussian distributions to the temperature forecasts, the histogram is much closer to uniformity. For wind speed, we show results for two EMOS variants, using truncated Gaussian (gEMOS) and truncated logistic (logEMOS) distributions, respectively. Both have similar results and improve the level of calibration significantly. More details about the composition and interpretation of these histograms can be found in Section 4.1.

Likewise, we can assess the benefit of statistical post-processing by looking at proper scores. Table 3.1 lists the values of the CRPS and the root-mean-square error (RMSE) for the MOGREPS-UK and EMOS temperature forecasts, averaged over all model runs initialised at 15 UTC, as well as all locations and lead times. While the CRPS looks at the whole probability distribution and assesses both calibration and sharpness,

[1] Visualisation for the histograms in this thesis is taken from Barnes, Brierley, and Chandler (2019).

Table 3.1: Continuous ranked probability score and root-mean-square error (in °C) of raw and post-processed MOGREPS-UK temperature forecasts, averaged over all forecast cases from the 15 UTC run in the test set. All pairwise score differences are statistically significant at $\alpha = 0.01$.

	CRPS	RMSE
Raw ensemble	0.734	1.253
EMOS	0.596	1.136

Table 3.2: Continuous ranked probability score and root-mean-square error (in knots) of raw and post-processed MOGREPS-UK wind speed forecasts, averaged over all forecast cases from the 15 UTC run in the test set. All pairwise score differences are statistically significant at $\alpha = 0.01$.

	CRPS	RMSE
Raw ensemble	2.116	3.670
gEMOS	1.618	3.056
logEMOS	1.622	3.070

the RMSE is a measure for the accuracy of the mean forecast. Again, more details are given in Section 4.1. We apply a permutation test (Heinrich et al., 2020) to the scores to test if the mean differences are significantly different. The corresponding scores for MOGREPS-UK wind speed forecasts and the two EMOS variants applied are shown in Table 3.2.

For both weather variables, EMOS manages to considerably improve the deterministic and probabilistic forecast skill, as compared to the original ensemble. The two EMOS approaches gEMOS and logEMOS perform on a similar level, with gEMOS receiving marginally lower scores. However, the logEMOS forecasts seem to be slightly better calibrated (Figure 3.2b). In the following, we concentrate on forecasts post-processed with the gEMOS method.

At this stage in the post-processing process, after the numerical model has run and EMOS (or a similar method) has been applied, the forecasts are issued for the full range of lead times. Typically, there are no further changes to this trajectory and at some point it is replaced by or blended with forecasts from newer model runs. With RAFT, we can repeatedly correct errors that become apparent as parts of the trajectory realise. First, we concentrate on improving the skill of the EMOS mean as a deterministic forecast.

RAFT is based on the notion that errors at forecast lead times are often correlated, as long as those lead times are sufficiently close to each other. We define the observed error of a deterministic forecast $m_{t,l}$, initialised at a time t with lead time l as the distance to the observation valid at the same time, y_{t+l}:

$$e_{t,l} = y_{t+l} - m_{t,l}. \tag{3.1}$$

Here, the deterministic forecast m is the mean of the EMOS predictive distribution and lead times l are measured in hours. For wind speed, the mean has to be calculated from the location and scale parameters first. Figure 3.3 illustrates the correlation matrix of the forecast errors over the training period at two locations in the UK. Neighboring lead times are highly correlated and the correlation becomes weaker as the distance between lead times increases. Therefore, the observed error at one particular lead time can be a reliable predictor for the expected error at a later lead time. We define the RAFT *adjustment period* as the period preceding a lead time l, where an observed error e_{t,l^*} at time $t + l^*$ with $l^* < l$ provides information about the error $e_{t,l}$. For temperature, the length of the adjustment period seems to vary as the predictability changes with the diurnal cycle (Figure 3.3a), while it seems more consistent for wind speed (Figure 3.3b).

We connect the predicted to the observed error using linear regression and define the RAFT model for the estimated error $\hat{e}_{t,l}$ at a future lead time l as

$$\hat{e}_{t,l} = \hat{\alpha} + \hat{\beta} \cdot e_{t,l^*} + \varepsilon, \tag{3.2}$$

with ε being normally distributed with mean zero. The assumption of normally distributed residuals holds for both the temperature and wind speed errors in our estimation dataset, as confirmed by Q-Q and residual plots (not shown). We estimate the regression coefficients $\hat{\alpha}$ and $\hat{\beta}$ using the least squares method for every possible combination of $l = 3, \ldots, 36$ hours and $l^* = 1, \ldots, 34$ hours. In order to make RAFT operationally viable, we allow for one hour to process observations and start with the first adjustment at time $t + 3$, using the observed error at $t + 1$.

From the coefficient $\hat{\beta}$, the length of the adjustment period, p, can now be determined for every l. If e_{t,l^*} is a useful predictor for $e_{t,l}$, then $\hat{\beta}$ in (3.2) will be significantly greater than zero for this combination of lead times. Below, we describe the algorithm to find the optimal length of p for lead time l, again using only the estimation period data. We here restrict p to be smaller or equal to 22 hours (with one additional hour reserved for processing observations), but it is possible to allow longer adjustment periods, especially in the temperature case, where predictability depends considerably on the diurnal cycle. The RAFT algorithm is defined in Paper II as follows:

1. Estimate the regression coefficients in (3.2) for all predictors e_{t,l^*} with l^* in $[l-23; l-2]$. If this results in any negative lead time values, we add 24 hours to l^*, so that lead time 23 is followed by lead time $0, 1, 2, \ldots$.

2. a) Find the earliest l^* in $[l-11; l-2]$, such that the coefficient $\hat{\beta}$ is significantly different from zero at the 10% level for each lead time $l^*+1, \ldots, l-2$. Denote the result as l_p.

   b) If there is no result in the previous step, find the earliest l^* in $[l-19; l-12]$, such that $\hat{\beta}$ is significantly different from zero at the 5% level for each lead time $l^*+1, \ldots, l-12$. Denote the result as l_p.

   c) If there is no result in the previous step, find the earliest l^* in $[l-23; l-20]$, such that $\hat{\beta}$ is significantly different from zero at the 1% level for each lead time $l^*+1, \ldots, l-20$. Denote the result as l_p.

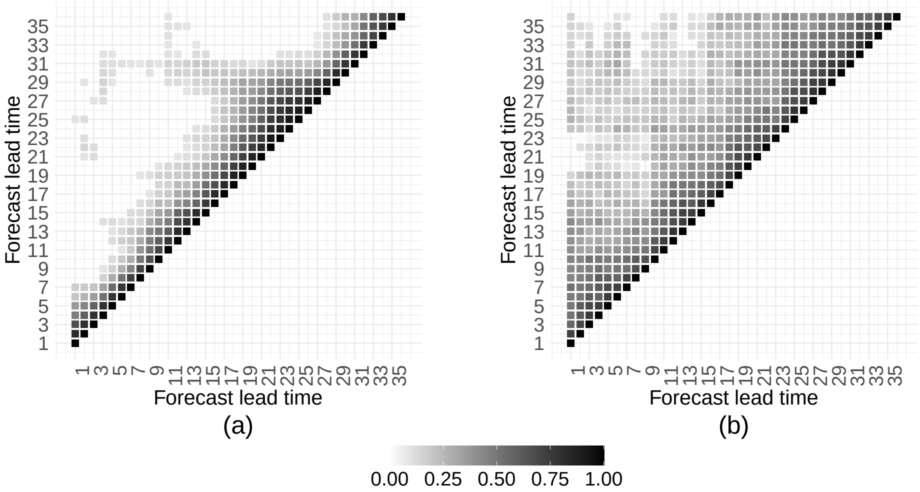


Figure 3.3: (a) Figure 4a from Paper I. Empirical correlation coefficient of the temperature mean forecast error for every lead time combination of the 03 UTC model run at Heathrow Airport during the training period. Only correlations significant at the 10% level are shown. (b) Figure 4a from Paper II. Empirical correlation coefficient of the wind speed mean forecast error for every lead time combination of the 15 UTC model run at The Cairnwell during the training period. Only correlations significant at the 10% level are shown.

3. After running the first two steps for all lead times, determine the length of the adjustment period p:

   a) If Step 2 has yielded a result for l_p, set $p = l - l_p$.

   b) If Step 2 has not yielded a result for l_p, set p equal to the average of the adjustment period length values for the neighbouring lead times $l - 1$ and $l + 1$.

   c) If there is still no valid value for p, set it to $p = 22$ hours. This corresponds to the longest possible adjustment period.

By letting the significance levels vary, we ensure that the adjustment period is as long as possible, while at the same time avoiding unreasonably long periods with negligible adjustments. In the few instances in this study where the algorithm was not able to identify a suitable adjustment period, this seemed due to significant, if small, error correlations for a large number of lead times. Therefore, it was decided to then set the adjustment period to the maximum value, which seemed to be justified, as tests showed no deterioration in forecast skill, even if the adjustment period is very long. This algorithm can be understood as a template that has been customised to the MOGREPS-UK ensemble. For other data sets, a different approach might be more suitable.

Finally, we collect the respective adjustment period lengths for each lead time l and the corresponding RAFT coefficients in (3.2) for every l^* in $[l - p; l - 2]$. All locations and model initialisation times are treated separately to account for location- and run-specific patterns. As can be seen from Figures 4b in Paper I and 2b in Paper II, the length of the adjustment period is not necessarily consistent between consecutive lead times. A smoother result might be achieved by pooling several lead times, however we did not see any jumpiness in the actual forecast skill of the RAFT updates. Also, varying predictability due to the diurnal cycle can more effectively be considered if adjustment period lengths are lead-time-specific.

Once we have obtained the coefficients, RAFT can be applied to online NWP model runs by plugging the latest observed errors into the RAFT model (3.2). The first adjustment of each model run is carried out at $t+2$ for all forecasts in the trajectory for which the respective adjustment period extends back to $t + 1$, the time the observation used in this step was recorded. From this point onward, hourly adjustments are applied until $t + 35$; Figure 3.4 illustrates how RAFT works in the context of a typical forecast cycle. The first two lead times can not be updated, as there are no forecast data from

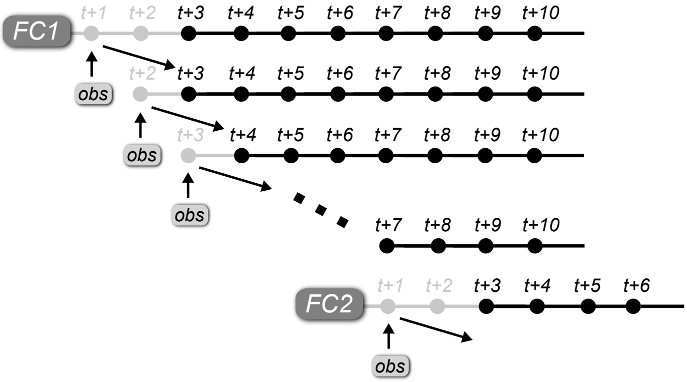


Figure 3.4: Figure 5 from Paper I. Diagram of a forecast cycle for an hourly forecast issued every 6 hours, with RAFT applied as new observations become available. Forecasts in grey are only used as predictors by means of their observed error and are not adjusted themselves.

the current run available. In order to somewhat bridge this gap, we use forecasts from the run initialised 24 hours earlier that are valid at the same time of day. However, this results in only a small improvement in forecast skill.

Formally, the predicted error $\hat{e}_{t,l}$ for a model run t and lead time l is calculated using the RAFT coefficients $\hat{\alpha}$ and $\hat{\beta}$ and the observed error $e_{t,l-k}$, where $2 \leq k \leq p$:

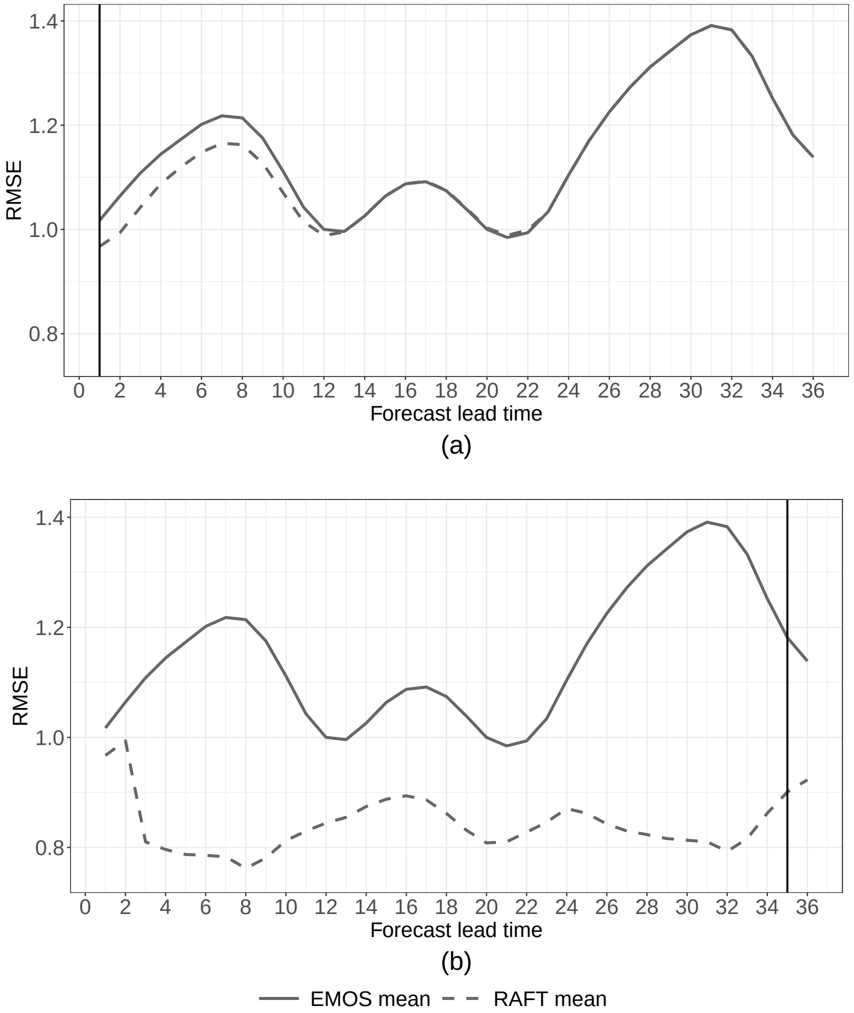
$$\hat{e}_{t,l} = \hat{\alpha} + \hat{\beta} \cdot e_{t,l-k}. \tag{3.3}$$

The estimated error is then added to the original EMOS mean forecast for lead time l to create the RAFT-adjusted forecast $\hat{m}_{t,l}$:

$$\hat{m}_{t,l} = m_{t,l} + \hat{e}_{t,l}. \tag{3.4}$$

It is possible for negative wind speeds to occur in this process, which we then set to zero. While here only the mean forecast is adjusted, it can now be plugged into the EMOS distribution in order to generate probabilistic forecasts. For wind speed, we first need to compute the location parameter, which we have to do numerically using the updated mean and the scale parameter.

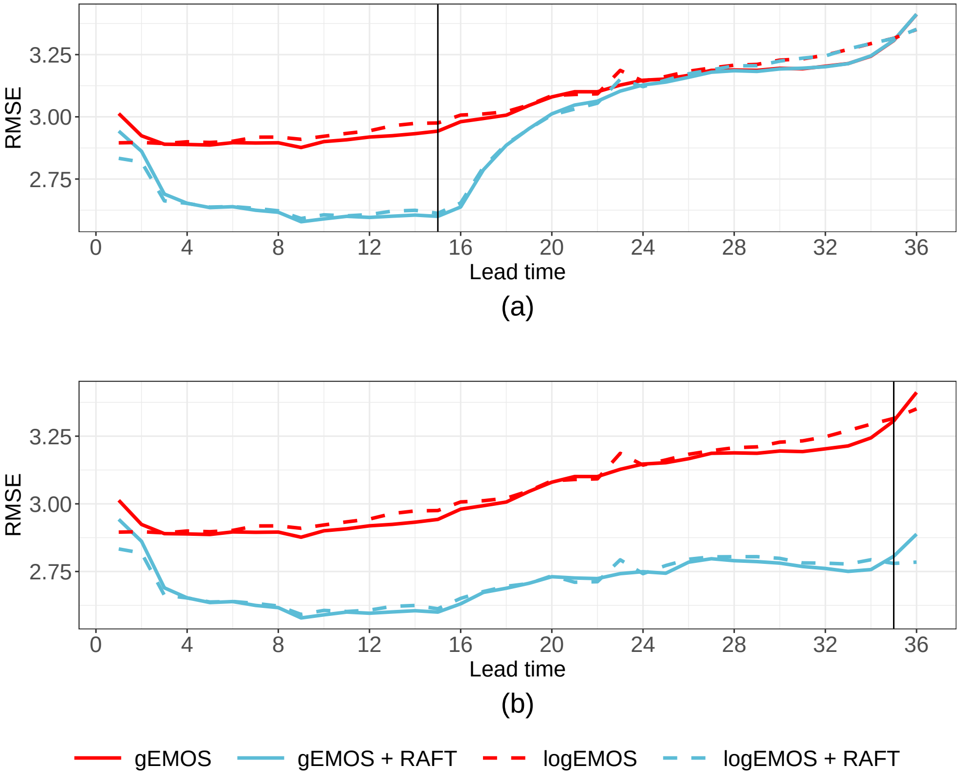
The deterministic and probabilistic forecast skill of the combined RAFT/EMOS forecasts is evaluated over the test period with the tools described in Chapter 4. To assess the accuracy of the mean forecast, we again use the RMSE. Figure 3.5 shows the average temperature RMSE over all sites and dates in the test period for the 21 UTC model run. The solid line is the RMSE of the EMOS mean over lead time and the dashed line the score for the RAFT-adjusted mean. In the top plot, RAFT is only applied once at $t+1$, using the error of the $t+24$ forecast from the previous day's run. Even with this limited additional information, there is a reduction in the RMSE for the next twelve hours.



(a)

(b)

Figure 3.5: Figure 9 from Paper I. (a) RMSE of the EMOS and RAFT mean temperature forecasts over lead time. The scores are averaged over all dates and locations in the test period for model runs initialised at 21 UTC. RAFT error corrections are carried out only once at lead time $t+1$. (b) is as (a), but RAFT is carried out for all lead times until the end of the trajectory.


(a)

(b)

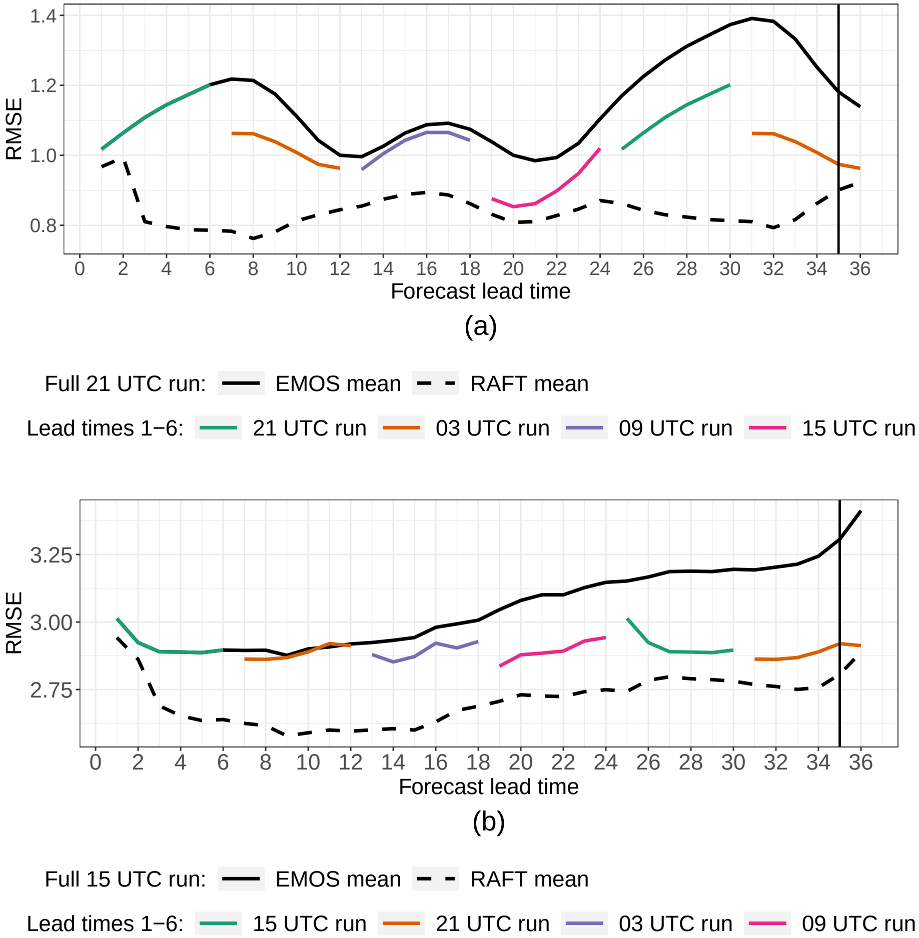
Figure 3.6: Figure 4 from Paper II. RMSE over lead time for gEMOS (red solid line) and logEMOS (red dashed line) mean wind speed forecasts, as well as their RAFT adjustments (blue solid and blue dashed lines, respectively). The scores are averaged over all dates and locations in the test period for model runs initialised at 15 UTC. (a) RAFT is only carried out until the adjustment at $t + 15$. (b) RAFT is carried out until its last iteration at $t + 35$.

The bottom plot shows the same data, but RAFT has now been applied hourly until the end of the trajectory, i.e., all RAFT adjustments have been made with the observations recorded two hours earlier, resulting in the maximum achievable forecast skill. There is a significant improvement of the deterministic skill across all lead times, but especially after $t + 3$, when we start using the forecasts from the current forecast run instead of the previous day's. It is noticeable that the RMSE of the RAFT mean is less subject to diurnal variations and the deterioration in skill due to increasing lead times completely disappears. The most substantial difference in forecast skill occurs during the last twelve hours of the trajectory.

In Figure 3.6, the same scores are shown for EMOS and RAFT wind speed forecasts. As noted before, the forecast skill here does not vary with the diurnal cycle. Figure 3.6a shows only the RAFT adjustments that have been made until $t + 15$, meaning that all forecasts to the left of the vertical line have been adjusted one hour before, and the ones to the right with the error information recorded at lead time $t + 14$. At this point, RAFT provides a skill increase for the next 10 hours, compared to the unadjusted EMOS forecasts. In Figure 3.6b, we see the wind speed equivalent to Figure 3.5b; the difference in RMSE between EMOS and RAFT is here quite consistent and increases slightly with lead time. Again, there is no significant difference between the two EMOS versions, gEMOS and logEMOS.

Both Figure 3.5b and Figure 3.6b illustrate one of RAFT's main benefits. We can compare the forecast skill at two lead times corresponding to the same time of day, for example $t + 2$ and $t + 26$. Usually, forecasts at $t + 26$ would have lower skill, as they lie in the tail of the trajectory. With RAFT, however, they are now more skilful than the forecasts at $t + 2$, as we have to rely on data from older forecast runs to adjust the first few hours of the trajectory, which is less effective. Consequently, it is reasonable to prefer the RAFT-adjusted forecasts at lead time $t + 26$ from the previous day's run to the $t + 2$ forecasts from the current one. Contrary to how NWP models currently operate, the newest model run has, for a brief time at the beginning of a trajectory, less forecast skill than forecasts from an older run. This time period is prolonged considerably when we take into account the computation time of the NWP model, which is disregarded in this study.

To further highlight the additional predictive skill provided by RAFT, we compare the final RAFT adjustments to the EMOS forecast from the most recent model run for any particular lead time (Figure 3.7). These forecasts have lead times of one to six hours and are therefore a fairer comparison to the final RAFT adjustments, which were created from information that was available two hours before the validity time. The top plot shows the same temperature scores as in Figure 3.5b, with the coloured lines being the RMSE values of the first forecasts from each model run. Again, the diurnal variation in predictability is quite noticeable, as is the gap in forecast skill between EMOS means at small and large lead times. The RAFT forecasts, however, are continuously performing better than EMOS, independent of lead time. Only the first 15 UTC run predictions come close to the RAFT-adjusted forecasts from the 21 UTC run during the period in the late afternoon where predictability increases, similar to the behaviour seen in Figure 8 in Paper I. The bottom plot shows the corresponding scores for wind speed forecasts. As expected, there is no significant diurnal cycle and the first forecasts from each run perform on a comparable level. During the first 18 hours of the trajectory, the RMSE of the RAFT mean is substantially lower than that of the newest EMOS predictions. However, as the skill of the trajectory deteriorates, the gap decreases. It should be kept in mind, though, that by this time the newer



(a)

(b)

Figure 3.7: (a) As Figure 3.5b, with added RMSE scores of the most current EMOS mean forecasts for each lead time, i.e., the first 6 hours of the most recent model run. (b) As Figure 3.6b, with added RMSE scores of the most current EMOS mean forecasts for each lead time, i.e., the first 6 hours of the most recent model run.

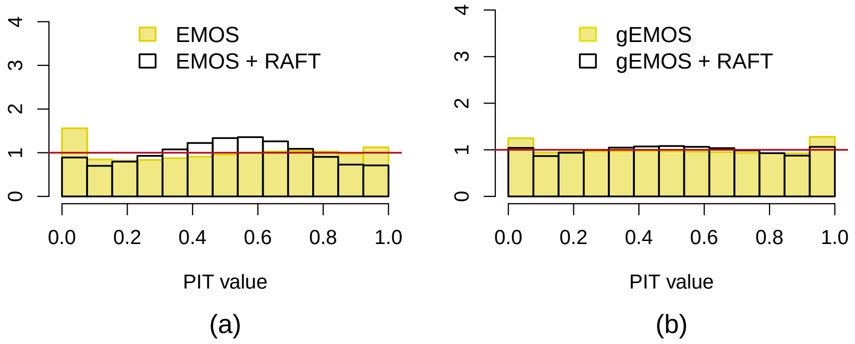


Figure 3.8: (a) PIT histograms of EMOS and RAFT mean temperature forecasts, aggregated over all sites and dates from model runs initialised at 15 UTC in the test data set. RAFT forecasts are taken from the final adjustment for each lead time. (b) PIT histograms of gEMOS and RAFT mean wind speed forecasts, aggregated over all sites and dates from model runs initialised at 15 UTC in the test data set. RAFT forecasts are taken from the final adjustment for each lead time.

EMOS runs themselves have been adjusted so that they are also now on a similar skill level as the first part of the RAFT mean curve in this figure.

The question arises if the combination of the RAFT-adjusted mean and the original EMOS variance constitutes a skilful probabilistic forecast, now that the deterministic skill has been improved. Thus, we look at the calibration of the RAFT/EMOS distribution via their probability integral transform (PIT) histograms in Figure 3.8. For both temperature and wind speed, the RAFT distributions now show a slight overdispersion, where they were a little underdispersed before. This is not surprising, as the spread includes some uncertainty that can be associated with the mean forecast and that has now been reduced by updating the mean. Although the sign of the miscalibration has changed, the coverage of the prediction interval shows that the RAFT/EMOS distributions are closer to perfect calibration than the original EMOS distributions.

3.2 RAFT for ensemble members

While adjusting only the mean forecast still results in calibrated distributions for MOGREPS-UK, this is not guaranteed for any type of ensemble. In our example, the EMOS forecast were slightly underdispersed and benefited from increasing the deterministic skill while keeping the EMOS variance unchanged. We now want to adapt RAFT in such a way that it works for NWP ensembles with different error profiles and adjusts the variance as well as the mean. To achieve this, we show how the RAFT technique for mean forecasts described in the previous section can be applied to individual ensemble forecasts. We call the mean forecast version RAFT$_{\mathrm{m}}$ and the ensemble member version RAFT$_{\mathrm{ens}}$. Results in this section are shown for wind speed forecasts only.

First, we sample ensemble members from the EMOS predictive distributions to obtain a calibrated ensemble with the same number of members as the original MOGREPS-UK. Following the recommendation by Schefzik, Thorarinsdottir, and Gneiting (2013), we use twelve equidistant quantiles. Similar to (3.1), the forecast error for the ith ensemble member forecast $x_{t,l}^{(i)}$ from the model run initialised at time t and valid at lead time l is defined as

$$e_{t,l}^{(i)} = y_{t+l} - x_{t,l}^{(i)}, \quad i = 1, \ldots, 12. \tag{3.5}$$

Then we proceed as before by investigating the relationship between the errors at different lead times over the estimation period. We look at every ensemble member individually and therefore do not treat them as exchangeable any more. The correlation matrix for one ensemble member of the 15 UTC run at The Cairnwell, Scotland is shown in Figure 3.9a; compared to Figure 3.3b, there seems to be less long-range correlation between earlier and later lead times. Figure 3.9b shows the corresponding adjustment periods for each lead time, determined using the algorithm described in Section 3.1. There is some jumpiness as to the length of the adjustment period at consecutive lead times, which is not present in the RAFT$_{\mathrm{m}}$ version (Figure 2b of Paper II). For larger ensembles than MOGREPS-UK, such as the 51-member ECMWF medium-range ensemble (Buizza and Richardson, 2017), it might be beneficial to pool ensemble members to achieve a smoother result and reduce the computational burden.

After determining the length of the adjustment period and the RAFT coefficients for every ensemble member, initialisation time, location and lead time combination, we

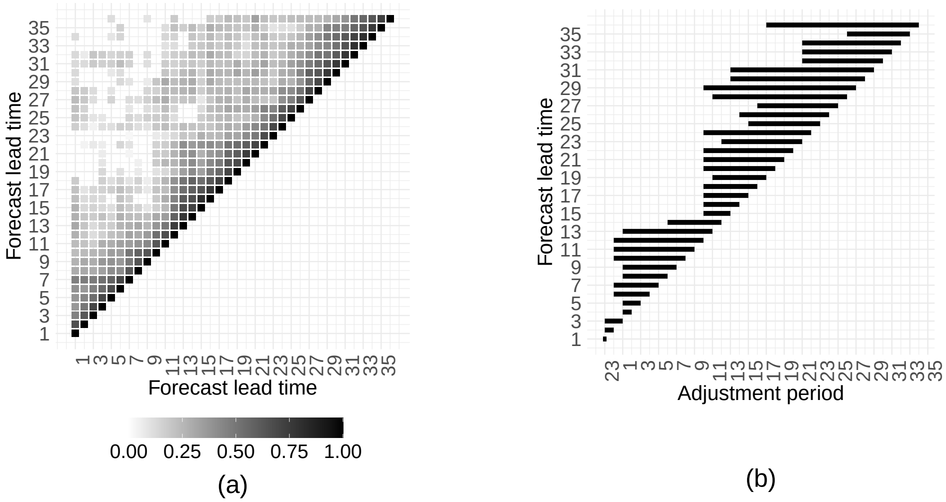


Figure 3.9: (a) Correlation matrix of an EMOS ensemble member for the 15 UTC run at The Cairnwell, with only correlations significant at the 10% level shown. (b) Corresponding RAFT$_{\text{ens}}$ adjustment period length for each lead time and the same ensemble member.

proceed with the online RAFT adjustments. The observed errors $e^{(i)}_{t,l-k}$ with $2 \leq k \leq p$ and the coefficients are plugged into the RAFT model

$$\hat{e}^{(i)}_{t,l} = \hat{\alpha} + \hat{\beta} \cdot e^{(i)}_{t,l-k} \tag{3.6}$$

and the predictive error $\hat{e}^{(i)}_{t,l}$ is added to the respective ensemble forecast:

$$\hat{x}^{(i)}_{t,l} = x^{(i)}_{t,l} + \hat{e}^{(i)}_{t,l}. \tag{3.7}$$

At any given lead time, the set of adjusted ensemble members $\hat{x}^{(i)}_{t,l}$ constitutes an empirical probability distribution, whose mean and variance have both been corrected by incorporating the most recent error information.

Again, when assessing the probabilistic skill of RAFT$_{\text{ens}}$ and of the combined RAFT$_{\text{m}}$ and EMOS distributions, we should keep in mind that the results for the latter method can vary considerably, depending on the properties of the ensemble. Figure 3.10 shows the verification rank histograms of the RAFT$_{\text{ens}}$ forecasts when applied to the two EMOS wind speed variants. These histograms can be directly compared to the ones in Figure 3.8, which seem overall to be slightly better calibrated. The scores in Table 3.3 confirm this result; in terms of the CRPS, RAFT$_{\text{m}}$ performs better, likely due to being closer to perfect calibration. RAFT$_{\text{ens}}$ has a slightly lower RMSE, but

Table 3.3: Continuous ranked probability score and root-mean-square error (in knots) of different combinations of EMOS and RAFT methods, averaged over all forecast cases from the 15 UTC run in the test set. The RAFT forecasts are taken from the final adjustments for each lead time. All pairwise score differences are statistically significant at $\alpha = 0.05$, apart from the RMSE differences between identical combinations using gEMOS and logEMOS.

	CRPS	RMSE
gEMOS + RAFT$_{\mathrm{m}}$	1.445	2.713
logEMOS + RAFT$_{\mathrm{m}}$	1.443	2.714
gEMOS + RAFT$_{\mathrm{ens}}$	1.483	2.708
logEMOS + RAFT$_{\mathrm{ens}}$	1.482	2.709

both methods improve the forecast skill considerably compared to the baseline EMOS forecasts in Table 3.2.

Figure 3.11 shows the mean CRPS at every RAFT iteration. As expected, the scores decrease as the forecast skill increases with each update. Throughout the process, RAFT$_{\mathrm{m}}$ has the lowest CRPS and the gap between the two methods widens. Although RAFT$_{\mathrm{ens}}$ exhibits the higher deterministic forecast skill, the CRPS here seems to reward the better calibrated RAFT$_{\mathrm{m}}$ forecasts. These results, together with the verification rank histogram in Figure 3.10, suggest that by applying RAFT$_{\mathrm{ens}}$, the predictive

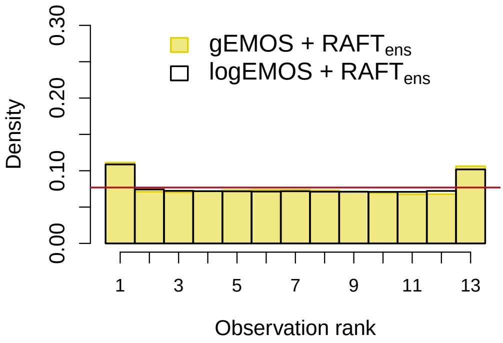


Figure 3.10: Verification rank histogram of the gEMOS and logEMOS + RAFT$_{\mathrm{ens}}$ forecasts, aggregated over all dates, locations and lead times in the evaluation set where the NWP model was initialised at 15 UTC. RAFT forecasts have been adjusted using the observations recorded 2 hours earlier.

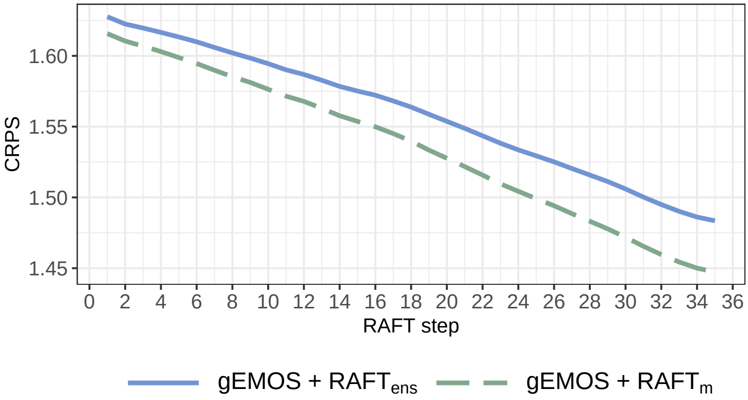


Figure 3.11: Figure 7a from Paper II. Mean CRPS for every step in the RAFT$_{\text{m}}$ and RAFT$_{\text{ens}}$ process. Scores are averaged over all lead times, sites and dates for the 15 UTC model initialisation time.

variance becomes too small at later RAFT iterations. This is indeed a danger with RAFT$_{\text{ens}}$: with consecutive updates, the forecasts become more and more certain while the spread decreases. As a result, later predictions may then be overconfident and show underdispersion. Although it did not present a large problem in our study, this feature should be kept in mind and thoroughly checked when applying RAFT$_{\text{ens}}$. Future research should include investigating ways to safeguard against underdispersion, such as in other member-by-member post-processing approaches (e.g., Van Schaeybroeck and Vannitsem, 2014).

To summarise, RAFT$_{\text{m}}$ and RAFT$_{\text{ens}}$ are largely comparable in terms of deterministic and probabilistic skill, with RAFT$_{\text{m}}$ performing marginally better overall. This can, however, be attributed to the properties of the ensemble and is not necessarily universal.

3.3 Order of operation for post-processing of multivariate forecasts

As discussed in Section 2.2, there is often a need for post-processed forecasts that are coherent in multiple dimensions, such as across locations, variables or lead times. Methods like ECC (Schefzik, Thorarinsdottir, and Gneiting, 2013) have been developed to work in combination with univariate post-processing methods to address this issue.

Table 3.4: Multivariate scores for different combinations of post-processing methods, averaged over all locations, lead times and dates in the test set, where the model was initialised at 15 UTC. RAFT forecasts are taken from the final adjustment for each lead time. The variogram score was calculated with order 0.5, as per the recommendation in Scheuerer and Hamill (2015). All score differences are significant at the 5% level.

	Energy score	Euclidean error	Variogram score
gEMOS + ECC	12.312	16.549	812
gEMOS + RAFT_m	11.943	15.045	899
gEMOS + RAFT_m + ECC	11.175	15.049	784
gEMOS + ECC + RAFT_ens	11.164	15.024	786

In Paper II, we clarify how the new RAFT method can be used together with EMOS and ECC in multiple post-processing stages and investigate the optimal order of operation.

In this application, we are interested in preserving the multivariate structure between the 36 hourly lead times in order to obtain coherent forecast trajectories. We create two alternative post-processing chains, based on the two RAFT versions described in the previous sections: EMOS + RAFT_m + ECC and EMOS + ECC + RAFT_ens. In the first scenario, we apply RAFT to the EMOS mean, then sample from the RAFT/EMOS distribution and reorder the new ensemble members according to the dependency template gathered from the raw ensemble. This setup means that ECC has to be applied at every RAFT adjustment step, resulting in longer computation times. In the second scenario, we sample from the unadjusted EMOS distribution first, reorder the new ensemble members only once and then apply RAFT_ens.

For assessing the multivariate forecast skill, some of the tools mentioned in Section 4.3 are employed. In Table 3.4, the average energy score and variogram score are given, along with the Euclidean error, which measures the Euclidean distance of the spatial predictive median to the observation vector. The results for the deterministic skill transfer from the univariate to the multivariate setting, as the forecasts produced by RAFT_ens have both the lowest RMSE and Euclidean error. Combining ECC and any version of RAFT reduces the energy score substantially, as compared to only using one of these methods. We can attribute the fact that gEMOS + ECC + RAFT_ens has a better score than gEMOS + RAFT_m + ECC to the higher deterministic skill of the former, as the energy score is much more sensitive to the mean error than to misspecifications in calibration or correlation (Pinson and Tastu, 2013). In the context of the variogram score, which puts more weight on a correct correlation structure, both combinations perform on the same level, with gEMOS + RAFT_m + ECC being

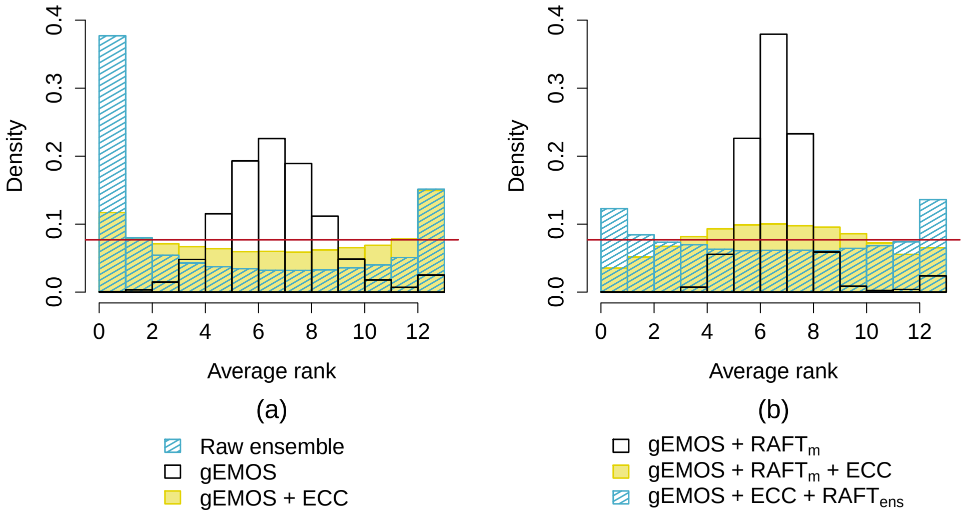


Figure 3.12: Figure 6 from Paper II. Average rank histograms for different combinations of post-processing methods. Data points are aggregated over all sites, model runs and lead times. All RAFT forecasts have been adjusted using the observation measured 2 hours earlier.

slightly better. This also means that the dependency template of the ensemble that was reintroduced by ECC is not being destroyed by applying RAFT$_{\text{ens}}$.

Finally, we are also interested in whether the multivariate forecasts produced by the two combinations of post-processing methods are calibrated. To this end, we look at the average rank and band-depth histograms. For the interpretation of these histograms, we refer to Figure 6.2 in Paper IV. Figures 3.12 and 3.13 illustrate how EMOS and ECC operate: the raw ensemble is underdispersed, but by applying EMOS to achieve better calibration, we lose all correlation between the lead times. ECC manages to restore the correlation structure, yet the forecasts are still somewhat underdispersed. As already seen with the univariate histograms, the predictive variances of gEMOS + ECC + RAFT$_{\text{ens}}$ are slightly too small and therefore underdispersive. The histograms for gEMOS + RAFT$_{\text{m}}$ + ECC are also not completely flat and show signs of a correlation structure that is too weak. This might be a remnant of the MOGREPS-UK ensemble not being able to specify the correlation perfectly in the first place and this flaw propagating through the post-processing chain.

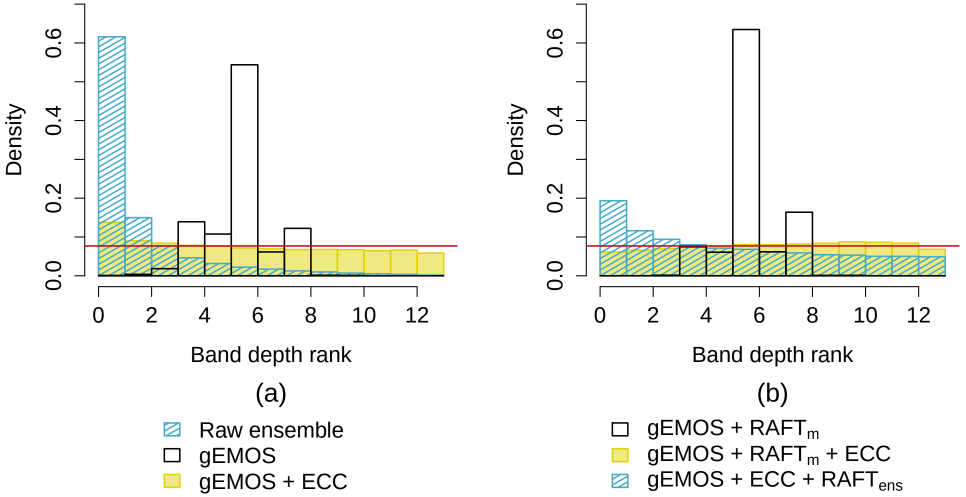


Figure 3.13: Band depth histogram for different combinations of post-processing methods. Data points are aggregated over all sites, model runs and lead times. All RAFT forecasts have been adjusted using the observation measured 2 hours earlier.

3.4 Forecast jumpiness and consistency

One major concern or difficulty in operational weather forecasting that has received increasing attention in recent years is forecast jumpiness. A forecast for a fixed time and location, say a particular event, will be updated several times in a setting similar to MOGREPS-UK, for example every time a new model run is initialised. It is reasonable to expect that the forecast accuracy improves with every update, however sometimes forecasts exhibit jumpy behaviour in that they do not converge towards the observation or switch between two different weather scenarios (Ehret, 2010). This “flip-flopping” can interfere with the decision-making process and may lead to a loss in confidence in the forecast provider.

As applying RAFT$_\mathrm{m}$ reduces the forecast error, we are also interested in whether this affects the forecast jumpiness or consistency. To this end, we look at the convergence index proposed by Ehret (2010) and compare it against the raw ensemble and the EMOS forecasts for surface temperature. Other, similar tools include the flip-flop index (Griffiths et al., 2018), inconsistency index (Zsoter, Buizza, and Richardson, 2009), forecast convergence score (Ruth et al., 2009) and divergence index (Richardson, Cloke, and Pappenberger, 2020). We refer to a number of predictions for the same validity time and location as a forecast sequence.

The convergence index combines the concepts of *divergence*, where the forecast error of a deterministic forecast sequence does not decrease, and *oscillation*, where the sign of the forecast error changes with consecutive updates. For every validity time, we collect forecasts from all model runs and count the number of divergences d and oscillations o in the forecast sequence, allowing for a tolerance of 1°C. The convergence index for a forecast sequence of length N is then defined as

$$\mathrm{conv} = \frac{\sum_{i=1}^{N-1} (d_i + o_i)}{2\,(N-1)}. \tag{3.8}$$

The version employed here is unweighted and uses an absolute tolerance rather than a value relative to the observation, as that is more appropriate for temperature. A convergence index of 1 denotes a forecast sequence that deteriorates at every step, while a perfectly convergent forecast sequence has a convergence index of 0. For more details, see Ehret (2010).

Figure 3.14 shows the station-wise means of the convergence index for the EMOS and EMOS + $\mathrm{RAFT_m}$ post-processed forecasts against the raw ensemble mean convergence indices. All values are close to 0 and the different forecasters therefore quite consistent. By applying EMOS, consistency is lost at about two thirds of the sites, as compared to the original ensemble. This is not surprising, as we calibrate each of the 36 lead times separately. At the final RAFT iteration, however, we have not only managed to compensate for this loss, but also improve on the raw ensemble's consistency at almost all sites. There are no sites where the jumpiness increases when RAFT is applied to the EMOS forecasts. Thus, adjusting forecast trajectories with RAFT benefits both forecast accuracy and consistency, making it more useful to customers in multiple aspects.

3.5 Seasonal temperature forecasts

RAFT can be universally applied to almost any forecasting scenario that involves two or more consecutive lead times. In Paper III, we test the effectiveness of RAFT in the context of a sub-seasonal to seasonal setting by means of the Met Office's GloSea5 ensemble (MacLachlan et al., 2015).

For this purpose, we make use of the GloSea5 hindcast data set, provided by the Copernicus Climate Change Service (C3S; Copernicus Climate Change Service, 2020).

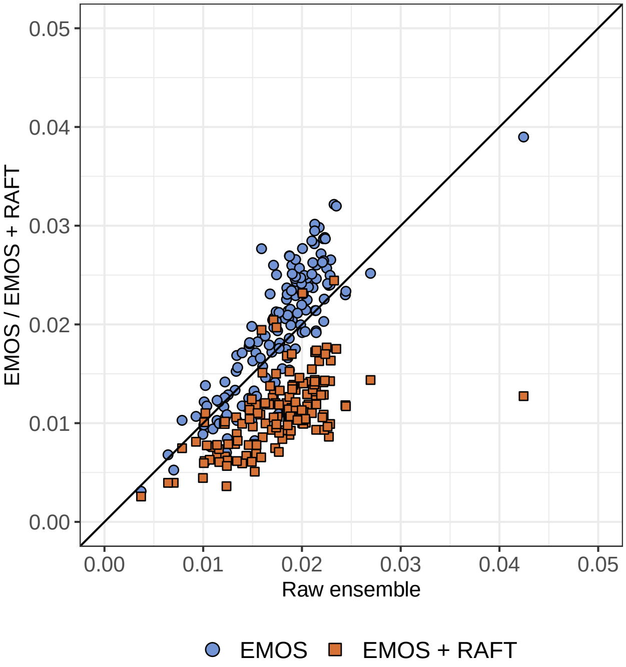


Figure 3.14: Convergence index for the EMOS and EMOS + RAFT$_m$ mean temperature forecasts as a function of the ensemble mean forecast convergence index, averaged for every observation station over all dates in the test data set. RAFT scores are taken from the final adjustment for each lead time.

This data set contains seasonal forecasts made by the same model version for the years 1993 to 2015, with seven ensemble members initialised weekly. We concentrate on forecasts for spring and summer surface temperature anomalies in Europe, covering the time period from early May until early September. As mentioned in Section 1.2, the ensemble members of three consecutive weeks are collected and treated as a lagged 21-member ensemble. The three sets of ensemble members have been initialised at different times and therefore from different analyses, with members from the newest run being on average the most skilful (Doblas-Reyes et al., 2013a; Doblas-Reyes et al., 2013b). If RAFT can manage to compensate for this discrepancy, the forecasts will be more accurate overall. To calibrate and verify GloSea5, we use ERA5 reanalyses (Copernicus Climate Change Service, 2017), regridded from their native grid resolution of 0.28° x 0.28° to the GloSea5 0.8° x 0.5° grid. Anomalies are calculated as the weekly/monthly

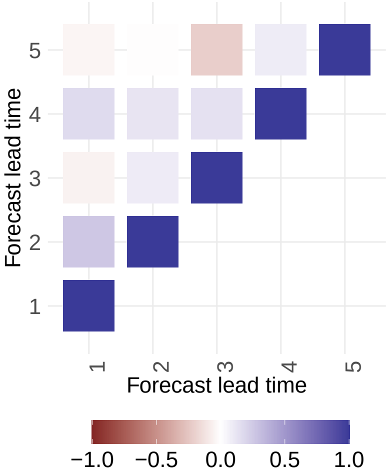


Figure 3.15: Correlations between forecast anomaly errors of the ensemble mean at monthly lead times for the model run initialised on 1 May. All correlations are significant at the 10% level.

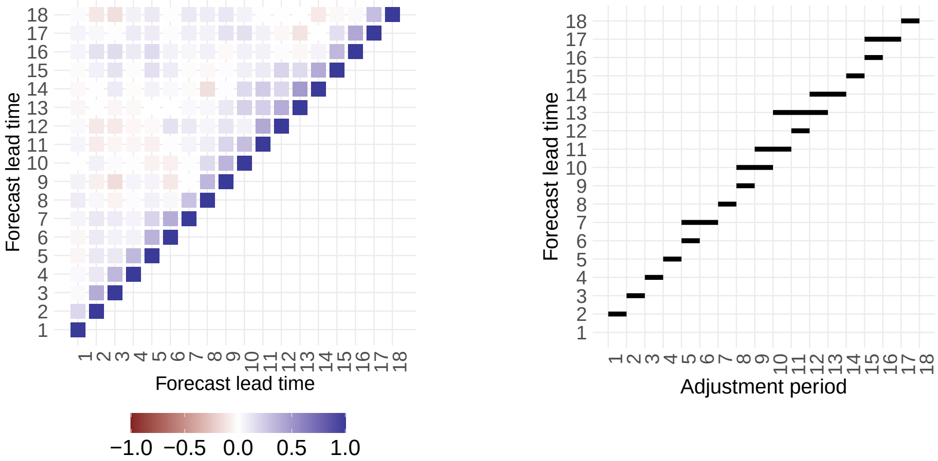
deviations of the ensemble mean and the reanalysis from their respective grid-point-wise temperature average over the same time period.

In a first step, we investigate the relationship between forecast lead times for monthly average forecasts, as these are the most widely used in terms of seasonal predictions. Because we are working with a limited data set, only five realisation times per year, we use data from all 23 years for this analysis and do not longer treat locations separately, but rather group all land grid points together. The forecast error $e_{t,l}$ of the ensemble mean $\bar{x}_{t,l}$ valid for lead time l and initialised at time t is again defined as

$$e_{t,l} = y_{t+l} - \bar{x}_{t,l}, \tag{3.9}$$

where y_{t+l} is the corresponding reanalysis value. Figure 3.15 shows the empirical correlation matrix of ensemble mean errors at different lead times for the run initialised on 1 May. Although all correlations are significant, there is no clear and noticeable pattern. The correlation values are quite jumpy and actually change sign. Experiments with applying RAFT using the shortest possible adjustment period of one month show almost no gains in forecast skill, which leads us to assume that the correlation between neighboring lead times is spurious rather than genuine.

Therefore, we move to a sub-monthly scale and repeat the analysis above for weekly averaged anomaly forecast trajectories, as seen in Figure 3.16a. Here, the correlation



(a)

(b)

Figure 3.16: Figure 2 from Paper III. (a) Correlations between forecast anomaly errors of the ensemble mean at weekly lead times for the model run initialised on 1 May. All correlations are significant at the 10% level. (b) The resulting adjustment periods for each forecast lead time.

between lead times is more substantial, although it decreases after a couple of weeks, matching the pattern from Figure 3.15. As we treat all grid points simultaneously and only have a maximum of 18 lead times, the length of the adjustment period p is determined subjectively and not with the algorithm described in Section 3.1. In this way, we avoid unrealistically long lead times and sign changes in the correlation. The resulting adjustment periods are shown in Figure 3.16b for the run started on 1 May. We repeat the process for the other runs initialised on 9 May, 17 May, 25 May and 1 June.

Again, the RAFT model relates the expected forecast error at lead time l with the observed error at an earlier lead time l^* through linear regression:

$$\hat{e}_{t,l} = \hat{\alpha} + \hat{\beta} \cdot e_{t,l^*} + \varepsilon. \tag{3.10}$$

The random term ε is normally distributed with mean zero and the coefficients $\hat{\alpha}$ and $\hat{\beta}$ estimated for every lead time combination using a leave-one-out-cross-validation approach. This means that we obtain the coefficients for a particular year using the complete data set apart from forecasts from that year. Anomaly mean forecasts $\bar{x}_{l,l}$ are then adjusted using the observed error at l^* if $l - p \leq l^* \leq l - 1$:

$$\hat{\bar{x}}_{t,l} = \bar{x}_{t,l} + \hat{e}_{t,l}. \tag{3.11}$$

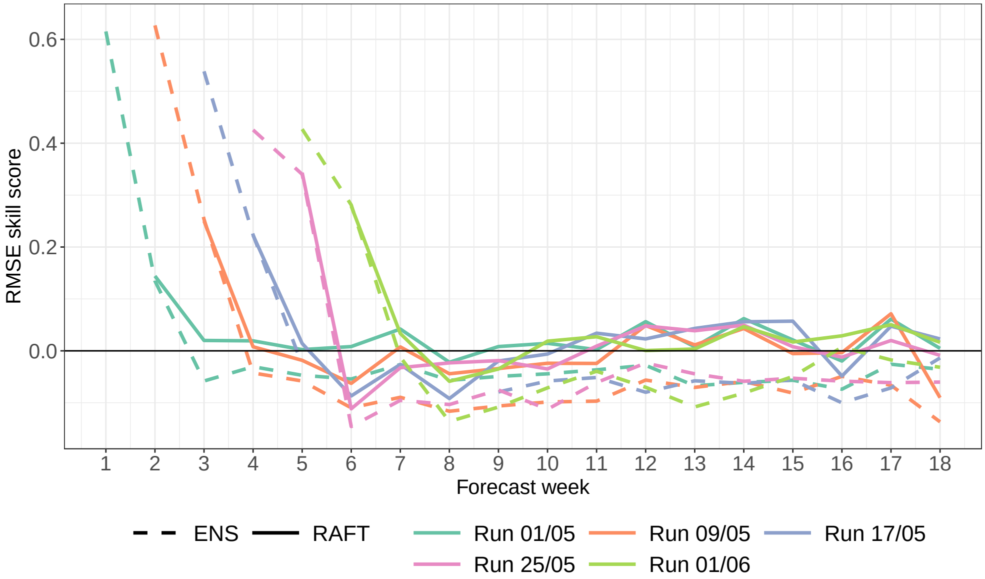


Figure 3.17: Figure 3 from Paper III. RMSE skill scores for five different runs of GloSea5 compared against ERA5 climatology. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015. The original GloSea5 forecasts are indicated with dashed lines, while RAFT forecasts updated one week prior to the realisation time are indicated by solid lines.

The goal of any long-range NWP forecast is to perform better than a climatological forecast, which we here compute from the ERA5 reanalyses. Therefore, we mainly look at the RMSE skill score (for details see Section 4.1) with the weekly local climatology as reference forecast. A perfect score would be 1, while a negative score means that the forecast has less skill than climatology. First, we evaluate the performance of the individual GloSea5 runs and their RAFT adjustments over lead time in Figure 3.17. For the first two weeks of a run, the raw ensemble has considerably more skill than climatology, but beyond week three, the ensemble's RMSE is about 5 to 15% worse than that of the climatology. RAFT forecasts from the final adjustment perform either slightly better or slightly worse than climatology, but always better than the raw ensemble.

As we are interested in the performance of a lagged ensemble, we also compute the unweighted average of all ensemble means available at any given lead time, for both GloSea5 and RAFT (Figure 3.18). In addition to this, the climatological forecast can also be updated with RAFT, using an adjustment lead time of one week. As long as new forecast runs are added to the lagged ensemble mean, the GloSea5 forecasts perform better than climatology, but drop to about the same level afterwards – a 5 to 15% improvement compared to the individual runs. Similarly, the combined RAFT forecast deteriorates after the last model run is added, but always has considerably more forecast skill than climatology. In the first few weeks, the combined RAFT mean performs worse than the lagged raw ensemble mean due to the stark discrepancy in predictive skill between weeks one and two (cf. Figure 3.17). As we can not adjust the first week of each forecast, the RAFT lagged mean here only includes forecasts from the second week of a model run onward. The skill of the adjusted climatological forecast is consistently about 5 to 10% higher than that of the unadjusted climatology and on par with the RAFT lagged mean. This illustrates once more how crucial recent observational information is and to what extent NWP becomes more challenging with increasing lead times. We could also try to adjust the lagged mean itself, but the

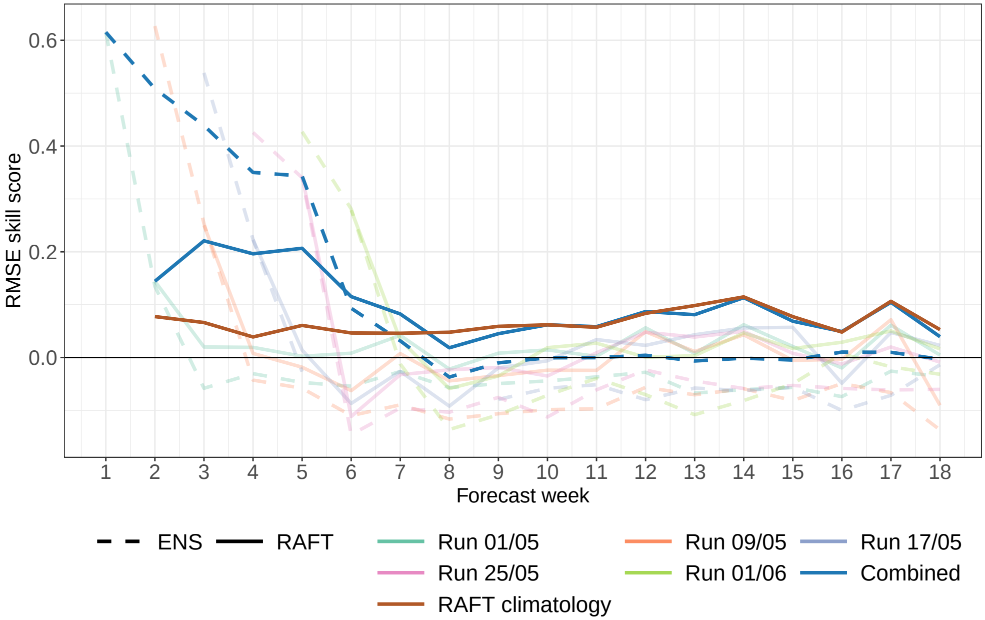


Figure 3.18: Figure 4 from Paper III. RMSE skill scores for a comparison against ERA5 climatology for the combination of all five forecast runs (blue dashed line), for RAFT-processed ERA5 climatology of adjustment lead time one week (brown solid line), and for a combination of the RAFT-processed ensemble means for all five forecast runs at an adjustment lead time of one week (blue solid line). For comparison, these skill scores are overlaid on the results shown in Figure 3.17.

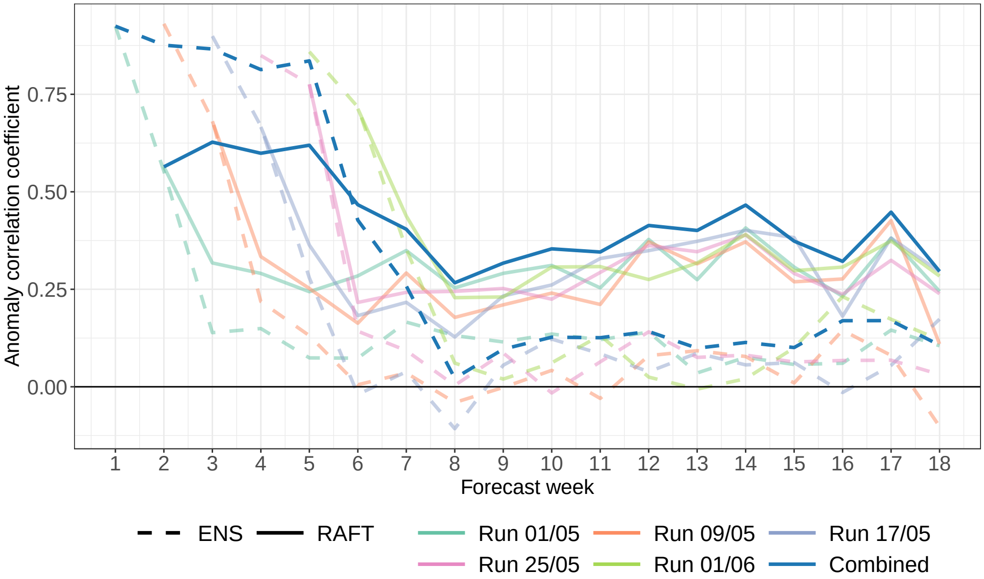


Figure 3.19: Anomaly correlation coefficient of different mean forecasts and the ERA5 reanalysis. Shown are the GloSea5 individual runs and their RAFT adjustments in the background and the combined GloSea5 mean forecast (dashed dark blue line), as well as the combined RAFT mean forecast (solid dark blue line). All RAFT forecasts have one week adjustment lead time.

primary goal is to correct systematic errors present in the individual model runs. Tests also show that the correlation pattern between lead times of the lagged ensemble mean is slightly weaker than seen in Figure 3.16, supporting the approach of applying RAFT first and aggregating afterwards.

A verification measure that is often used in conjunction with the RMSE is the anomaly correlation coefficient (ACC), the correlation between the predicted and the observed anomalies; more details can be found in Section 6.4 of Paper IV. The ACC assesses how well the forecasting system tracks the observations over time without regarding the bias. In Figure 3.19, we compare the ACC of the five individual GloSea5 runs with the respective RAFT adjustments. Also shown are the combined mean forecasts for the raw ensemble and RAFT. After week seven, when there is no additional data from new model runs, the ensemble correlation does not exceed 0.25 and the combined forecast performs mostly on the same level as the best individual model run. The correlation of the RAFT forecasts on the other hand always lies between 0.25 and

0.5 and the lagged RAFT mean outperforms the individual models at almost all lead times.

The question remains if we should always combine all forecasts that are available at a particular realisation time or if only certain runs should be selected. Therefore, we look at all ensemble forecasts for lead times greater than week five and at the RAFT forecasts that are issued in week five. Figure 3.20 shows different ways to combine the lagged forecasts, from only using the latest run initialised on 1 June to averaging the ensemble means of all five runs; we also include the RAFT-adjusted climatology forecast. For the first two weeks, there is indeed a benefit in combining only a few runs, whereas for the later part of the trajectory, all runs and the adjusted climatology should be combined. However, none of these forecasts consistently outperform climatology.

While using RAFT to incorporate additional error information into seasonal temperature trajectories provides a gain in forecast skill, it is only for a short time and does not go far beyond the medium range. For a useful application to seasonal forecasts, a

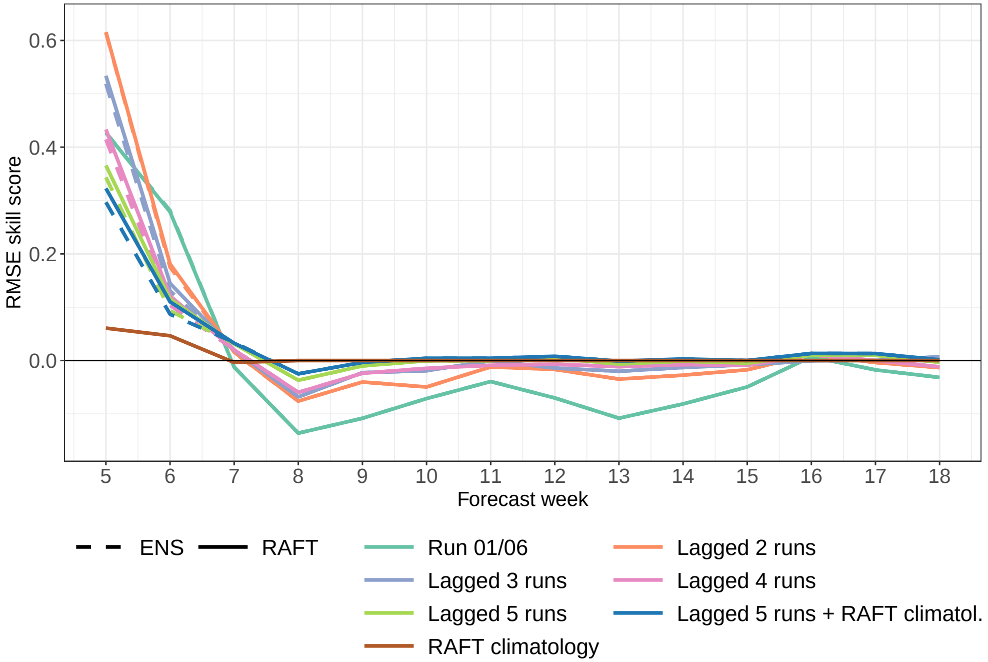


Figure 3.20: Figure 5 from Paper III. RMSE skill scores for various model combinations of forecasts issued in week five compared against ERA5 climatology. Each combination consists of the most recently available runs. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015.

substantial correlation between lead times further than a month apart would be needed. However, we see signs that creating lagged ensembles, especially if the ensemble members have been updated with RAFT, reduces forecast bias and increases the correlation between forecasts and reference. In order to improve the effectiveness of RAFT for seasonal predictions, future research could look into applying RAFT to forecasts of the underlying atmospheric circulation patterns that govern European seasonal variations. These patterns, such as the North Atlantic Oscillation, can be directly linked to surface weather variables, but are often easier to predict (Lledó et al., 2020).

Chapter 4

Forecast verification

In order to judge the goodness of a forecast and therefore the effectiveness of any NWP modelling or post-processing, verification methods are needed. They summarise information about calibration, sharpness or overall skill into one figure or number. Paper IV is an overview over the multitude of verification measures employed in weather forecasting and gives some advice how to best use them.

4.1 Univariate forecast verification

As stated in Section 1.3, there are two important properties to forecast skill and the goal of any post-processing should be to maximise the sharpness subject to calibration (Gneiting, Balabdaoui, and Raftery, 2007). Calibration is the statistical compatibility of forecasts and observations; ideally an observation would be indistinguishable from an ensemble member or a random sample from the predictive distribution. In practice, we can assess calibration with the probability integral transform (PIT) histogram (Dawid, 1984; Diebold, Gunther, and Tay, 1998).

The PIT histogram tests whether the forecasts and observations exhibit the same statistics over a long time period. To create the histogram, we compute the value of the predictive cumulative distribution functions (CDF) $F_1, \ldots, F_n$ at the corresponding observation $y_1, \ldots, y_n$ and then aggregate over a number of forecast cases n:

$$F_1\left(y_1\right), \ldots, F_n\left(y_n\right). \tag{4.1}$$

These PIT values are plotted in a histogram, which is uniform or flat if the forecasts are perfectly calibrated. Underdispersive forecasts can be recognised by a $\cup$-shaped histogram and overdispersive forecasts by a $\cap$ shape; if the histogram is monotonic (meaning a triangular shape), the forecasts are biased. Examples of these histograms can be found in Figure 6.1 of Paper IV. The verification rank histogram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand, Vautard, and Strauss, 1997) applies to ensemble forecasts and can be interpreted in the same manner as the PIT histogram. Given a calibrated ensemble, the observation has the same likelihood to occupy any rank in the joint set of ensemble members and observation. If we compute this rank for every forecast case and then plot the observation ranks as a histogram, we can again identify deficiencies in the calibration of the ensemble.

Note that a flat histogram is only a necessary condition for calibration, not a sufficient one (Hamill, 2001). It is therefore best to confirm the findings with other methods, like the coverage of the prediction interval spanned by the ensemble members. The closer to the nominal value of $(m-1)\,/\,(m+1)$ for an ensemble with m members, the better the calibration. For the simulation study in Section 6.3.2 of Paper IV, we created two forecasting scenarios that only differ by the amount of data points. Figures 6.4 and 6.5 in that section show that it can be hard to assess different forecasts if the sample size is not sufficiently large. A statistical significance test like the chi-squared test (Wilks, 2004; Wilks, 2011) can help in such situations to find out if a forecast is really calibrated.

For ranking competing forecasts and their accuracy, scoring rules (Gneiting and Raftery, 2007) are the best choice. A scoring rule S is a function $S(F, y)$ that assigns a numerical value to the skill of each probabilistic forecast F relative to the verifying observation y. The better the forecast matches the observation, the higher the accuracy and the lower the score. It is very important that the scores used are proper, which means that the expected score of the true forecast distribution should always be the lowest. This prevents hedging strategies and ensures that forecasters always issue the best prediction they can. Most scores are computed for every individual forecast case and then averaged over a large data set.

Two of the most popular scores are the continuous ranked probability score (CRPS) and the ignorance or logarithmic score. Both scores assess calibration and sharpness simultaneously and are thus used for verification as well as minimum score estimation (see also Section 2.1), where some parameters are estimated by minimising the score over a training period. The ignorance score (Good, 1952) is defined as the logarithm of the value of the predictive distribution, evaluated at the observation:

$$\mathrm{IGN}\left(F, y\right) = -\log\left(f\left(y\right)\right). \tag{4.2}$$

Here, f is the probability density function (PDF) of the distribution F. Optimising the ignorance score is thus equivalent to maximum likelihood estimation (Fisher, 1922). The ignorance score has the disadvantage that it is very sensible to outliers and does not apply directly to discrete ensemble forecasts.

For verification, the CRPS (Matheson and Winkler, 1976) is often preferred, as it is more robust and has several representations suitable for different kinds of forecasts

(Gneiting and Raftery, 2007; Gneiting and Ranjan, 2011; Hersbach, 2000; Laio and Tamea, 2007):

$$
\begin{aligned}
\mathrm{CRPS}\left(F, y\right) &= \mathbb{E}_F \left|X - y\right| - \frac{1}{2}\mathbb{E}_F\mathbb{E}_F \left|X - X'\right| && (4.3)\\
&= \int_{-\infty}^{\infty} \left(F\left(x\right) - \mathbb{1}\left\{y \leq x\right\}\right)^2 \,\mathrm{d}x && (4.4)\\
&= \int_0^1 \left(F^{-1}\left(\tau\right) - y\right) \cdot \left(\mathbb{1}\left\{y \leq F^{-1}\left(\tau\right)\right\} - \tau\right) \mathrm{d}\tau. && (4.5)
\end{aligned}
$$

Here, $\mathbb{1}\left\{\cdot\right\}$ denotes the indicator function and F^{-1} the quantile function of the predictive distribution F. The forms in (4.4) and (4.5) can be interpreted in terms of other proper scores, namely the Brier score (Brier, 1950) and the quantile score (Friederichs and Hense, 2007; Gneiting and Raftery, 2007), respectively. In (4.3), X and X' are independent random variables distributed according to F (Gneiting and Raftery, 2007). For ensemble forecasts, the first representation can be approximated (Grimit et al., 2006) by

$$
\mathrm{CRPS}\left(X_1, \ldots, X_m, y\right) = \frac{1}{m}\sum_{i=1}^{m} \left|X_i - y\right| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} \left|X_i - X_j\right|. \tag{4.6}
$$

As described in Section 2.1, we estimate the EMOS coefficients for post-processing the MOGREPS-UK forecasts in Paper I and Paper II by minimising the CRPS over the training period. There are closed forms of the CRPS for the Gaussian distribution in the temperature EMOS model in (2.1) (Gneiting et al., 2005), as well as for the two wind speed EMOS models using truncated Gaussian (2.4; Thorarinsdottir and Gneiting, 2010) and truncated logarithmic distributions (2.5; Scheuerer and Möller, 2015):

$$
\begin{aligned}
\mathrm{CRPS}\left(\mathcal{N}\left(\mu, \sigma^2\right), y\right) &= \sigma\left\{\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right] + 2\varphi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right\} && (4.7)\\
\mathrm{CRPS}\left(\mathcal{N}^{+}\left(\mu, \sigma^2\right), y\right) &= \sigma \cdot \Phi\left(\frac{\mu}{\sigma}\right)^{-2}\left[\frac{y-\mu}{\sigma} \cdot \Phi\left(\frac{\mu}{\sigma}\right)\left\{2\Phi\left(\frac{y-\mu}{\sigma}\right) + \Phi\left(\frac{\mu}{\sigma}\right)\right.\right. && (4.8)\\
&\qquad \left.\left. -2\right\} + 2\varphi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\Phi\left(\sqrt{2}\cdot\frac{\mu}{\sigma}\right)\right]\\
\mathrm{CRPS}\left(\mathcal{L}^{+}\left(\mu, s\right), y\right) &= \left(y - \mu\right) \cdot \left(\frac{2p_y - 1 - p_0}{1 - p_0}\right) + s \cdot \left[\log\left(1 - p_0\right)\right. && (4.9)\\
&\qquad \left. - \frac{1 + 2\log\left(1 - p_y\right) + 2p_y \mathrm{logit}\left(p_y\right)}{1 - p_0} - \frac{p_0^2 \log\left(p_0\right)}{\left(1 - p_0\right)^2}\right].
\end{aligned}
$$

The CDF and PDF of the standard normal distribution are denoted by Φ and φ, respectively; $p_0 = \Lambda\left(\mu s^{-1}\right)$ and $p_y = \Lambda\left(\left(y-\mu\right)/s\right)$ are values of the CDF of the standard logistic distribution, and $\operatorname{logit}\left(p\right) = p/\left(1-p\right)$ is the logit function.

For assessing deterministic forecasts, we use so-called scoring functions. Similar to scoring rules, a penalty is computed for every forecast x relative to the verifying observation y. It is essential that scoring functions are consistent for the target functional issued as a forecast (Gneiting, 2011). This means that e.g., if the issued forecast is the mean of a predictive distribution F, it should only be evaluated with an appropriate scoring function, such as the squared error

$$\mathrm{SE}\left(F, y\right) = \left(\operatorname{mean}\left(F\right) - y\right)^2 . \tag{4.10}$$

When averaging the squared error over a large number of forecast cases n, the root-mean-square error (RMSE) is often computed:

$$\mathrm{RMSE}\left(F, y\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathrm{SE}\left(F_i, y_i\right)}. \tag{4.11}$$

One scoring function consistent for the median of a forecast distribution is the absolute error

$$\mathrm{AE}\left(F, y\right) = \left|\operatorname{med}\left(F\right) - y\right| . \tag{4.12}$$

If applied to a deterministic forecast, the CRPS reduces to the mean absolute error.

In Paper IV, we show in a simulation study how the sample size can affect the outcome when comparing competing forecasts with proper scores. For this purpose, we test if scoring rules can identify the true forecast from a range of predictive distributions. Two sets of data are drawn from a true distribution, which is Gaussian in the first part of the study and a Gumbel distribution in the second. We use one data set as training data to estimate the moments of four competing forecast distributions (Gaussian, non-central t, log-normal and Gumbel) and the other to verify these forecasts with the absolute error, the squared error, the CRPS and the ignorance score. The fifth forecaster is the true distribution with the true parameters. This process is repeated twice, once with a total of 1000 forecast-observation pairs to test how the scores would rank the different forecasters in a somewhat realistic setting, and once with one million data points to find out the true order in which these forecasts should be ranked. Bootstrap intervals indicate if the difference between two mean scores is statistically significant at the 5% level.

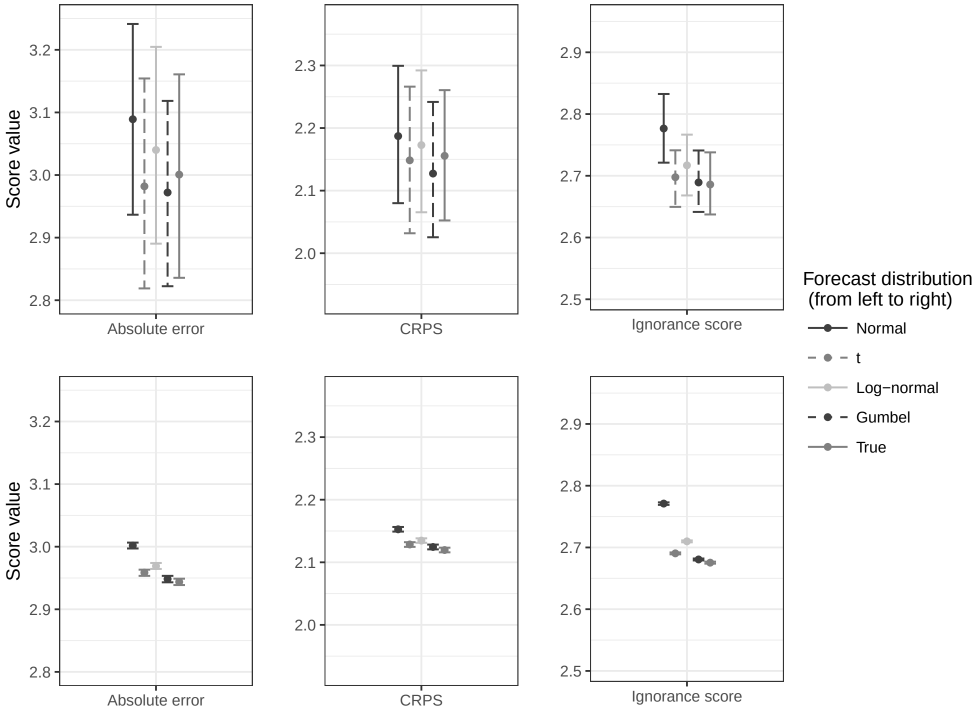


Figure 4.1: Figure 6.7 from Paper IV. Top row: Mean absolute error, CRPS and ignorance score, and the 95% bootstrap confidence interval for the five forecast distributions, if the true distribution is a Gumbel distribution. Scores are based on 1000 forecast-observation pairs. Bottom row: Same as above, but scores are based on 1 million forecast-observation pairs.

Figure 4.1 shows a plot from Paper IV with the ranking of the five forecasters according to different scoring rules if the truth is a Gumbel distribution. While the scores agree if the sample size is very large, only the ignorance score manages to identify the true distribution as the best for the smaller data set. The absolute error and the CRPS values of the non-central t-distribution are actually lower than the score for the true distribution. This confirms the findings in Gebetsberger et al. (2018) that the ignorance score is more sensitive to the shape of the distribution and outliers in general, while the CRPS puts more importance to the center of the distribution. From the results of this simulation study, we conclude that forecasts should be evaluated over as large a data set as possible and mean scores should be accompanied by confidence intervals.

Our assessment of whether two mean scores are significantly different is based on overlapping bootstrap intervals (e.g., Lahiri, 2003). It is also possible to use less empirical and more formal tests like the Diebold-Mariano test (Diebold and Mariano, 1995). Heinrich et al. (2020) propose a permutation test that has the advantage that the asymptotic variance does not have to be estimated. We use this permutation test in Paper II to compare the performance of different combinations of post-processing methods.

Instead of mean scores, the relative skill of two forecasts is often given by means of a skill score:

$$S_{\text{skill}} = \frac{\overline{S}_{\text{fc}} - \overline{S}_{\text{ref}}}{\overline{S}_{\text{perf}} - \overline{S}_{\text{ref}}}. \tag{4.13}$$

Here, the mean score $\overline{S}_{\text{fc}}$ of a particular forecast is compared to the mean score $\overline{S}_{\text{ref}}$ of a reference, often a climatological forecast, while $\overline{S}_{\text{perf}}$ is the score of a perfect forecast. A negative skill score means that the forecast performs worse than the reference, a skill score equal to 1 would indicate a perfect forecast. Any scoring rule can be chosen for S, but it should be noted that skill scores are not necessarily proper, even if computed from a proper scoring rule (Gneiting and Raftery, 2007; Murphy, 1973).

4.2 Verification of forecasts for extremes

Often, forecast providers and their customers are interested in predicting extreme events in order to assess risk, prevent damage to life and property and to mitigate impacts. There are several post-processing methods that focus on weather extremes; an overview can be found in Friederichs, Wahl, and Buschow (2018). In a similar fashion, special verification tools are required for highlighting a forecasting system's performance when it comes to extreme events.

The most important aspect of verifying forecasts for extremes is to not simply restrict the data set to a subset where the observations are extreme, e.g., above or below a certain threshold. Lerch et al. (2017) illustrate how such a strategy can lead to proper scoring rules becoming non-proper and therefore unable to assess forecasts in the correct manner. For example, a forecast constantly predicting an extreme value will be correct in 100% of the cases if only extreme observations are considered, but is very poor overall. This so-called “forecaster's dilemma” therefore requires us to use all available forecast-observation pairs and employ versions of the regular verification tools that are weighted towards the extreme part of the distribution.

Such scores include the conditional likelihood and censored likelihood scores, which are versions of the ignorance score (Diks, Panchenko, and van Dijk, 2011). A simulation study in Paper IV shows different options for using the threshold-weighted CRPS proposed by Gneiting and Ranjan (2011). This score is defined as

$$\mathrm{twCRPS}\,(F, y) = \int_{-\infty}^{\infty} w\,(z)\,(F\,(z) - \mathbb{1}\,\{y \leq z\})^2\,\mathrm{d}z, \tag{4.14}$$

where $w\,(z)$ is a non-negative weight function. Depending on the choice of $w\,(z)$, certain parts of the predictive distribution can be emphasised.

In a similar setting as in the previous section, we compare the threshold-weighted CRPS with three weighting functions,

$$\begin{aligned} w_1\,(y) &= \mathbb{1}\,\{y \geq u\}\,, \\ w_2\,(y) &= 1 + \mathbb{1}\,\{y \geq u\} \text{ and} \\ w_3\,(y) &= 1 + \mathbb{1}\,\{y \geq u\} \cdot u, \end{aligned}$$

to the unweighted CRPS and the non-proper CRPS with restricted observations, where the threshold u is selected as the 97.5% observed quantile. The true distribution is Gaussian and the four forecasters consist of two Gaussian and two Gumbel distributions. Figure 4.2 shows the mean scores and the respective 95% bootstrap intervals, apart from the first weight function, as these scores are equal to zero. It is obvious that the CRPS with restricted observations is not proper, as it disagrees with the other scores and assigns the lowest value to the Gumble distribution with fixed parameters, which has the heaviest tail of all forecast distributions. All other scores agree on the ranking of the forecasts, which makes the use of weighted scoring rules rather limited (Lerch et al., 2017), especially when the threshold quantile becomes large. Nevertheless, they can provide useful insight when it comes to the interpretation of forecast skill for extremes.

4.3 Multivariate forecast verification

Many forecasting applications require some multivariate consistency, whether spatially, temporally or between weather variables. In order to select the best forecasting methods that also incorporate dependencies, we need corresponding multivariate verification tools. The biggest challenge for these tools is that they have to work well on both small and large numbers of dimensions.

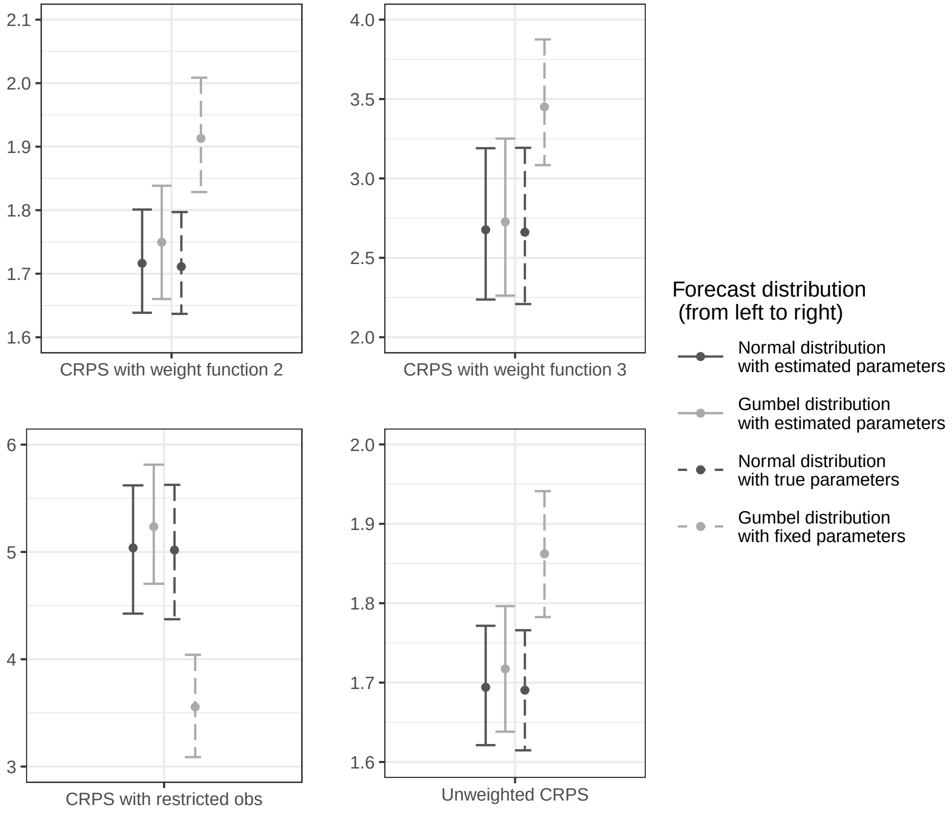


Figure 4.2: Figure 6.11 from Paper IV. Mean scores and 95% bootstrap confidence interval for the four versions of the CRPS. Top row: twCRPS with weight functions w_2 and w_3. Bottom row: CRPS restricted to observations above the threshold u and unweighted CRPS.

Multivariate calibration is probably the most difficult to assess, as the ranking of vectors across multiple dimensions is quite problematic. In Paper IV, we compare four histograms that all follow the same template. First, for each element in the set of ensemble member and observation vectors $S = \{X_1, \ldots, X_m, y\}$, each of length d, a so-called pre-rank is calculated. Then, the rank of the observation pre-rank is determined and plotted as a histogram. The four methods only differ in their approach to calculating the pre-rank.

The direct extension of the univariate verification rank histogram is the multivariate rank histogram (Gneiting et al., 2008), where the pre-rank is simply the sum of vectors that are smaller in every dimension. Unfortunately, this technique results in a large number of equal pre-ranks if the dimension is greater than two or three (Pinson and Girard, 2012) and the multivariate rank is heavily influenced by random decision, resulting in a potentially misleading flat histogram.

Thorarinsdottir, Scheuerer, and Heinz (2016) propose two alternatives, the average rank histogram and the band-depth histogram. For the average rank histogram, we first calculate the univariate ranks for each component of the vectors in S. The average over the univariate ranks of a vector is then used as the pre-rank in the second step of the scheme. The concept behind the band-depth histogram is slightly different, in that it is based on the notion that the components of the observation vector will on average lie in every band spanned by pairs of points from S with the same frequency if the ensemble is calibrated. In this context, ranks are not defined from small to large values, but from the center of a set of curves outward; thus the band-depth histogram can not be interpreted in the same way as the previous techniques.

To calculate the pre-rank for the minimum spanning tree histogram, we look at the minimum spanning tree (Smith and Hansen, 2004; Wilks, 2004) of the set S without the vector u, where $u = X_1, \dots, X_m, y$. The minimum spanning tree is a collection of pairs of points connected by segments in such a way that all points are used without closed loops, and the total length of the segments is minimised. Then, the pre-rank of a vector is the length of its minimum spanning tree.

These four tools, although similar in their approach, can behave very differently depending on the type of miscalibration in the data. Perfect calibration always results in a flat histogram, but diagnosing deficiencies can be difficult. In Figure 6.2 in Paper IV, we show the shapes of the rank histograms for different kinds of miscalibration in order to facilitate interpretation. Ideally, one should use several of these methods to confirm results, e.g., the average rank and the band-depth histograms like in Paper II.

Some of the proper scores described in Section 4.1 have multivariate extensions, while others can more or less be directly applied to multivariate predictive distributions, like the ignorance score. However, the outcome of multivariate post-processing methods is often in the form of ensembles and not full probability distributions. In these cases, the generalisation of the CRPS to multiple dimensions, the energy score (Gneiting and Raftery, 2007), is a popular choice. It is defined as

$$\mathrm{ES}\left(F, y\right) = \mathbb{E}_F \left\| X - y \right\| - \frac{1}{2} \mathbb{E}_F \mathbb{E}_F \left\| X - X' \right\|, \tag{4.15}$$

where X and X' are independent random vectors distributed with the multivariate predictive distribution F and y is the observation vector. The Euclidean norm is here denoted by $\|\cdot\|$. For an ensemble of vectors $X_1, \dots, X_m$, the energy score can be expressed as

$$\mathrm{ES}\left(X_1, \dots, X_m, y\right) = \frac{1}{m} \sum_{i=1}^{m} \|X_i - y\| - \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \|X_i - X_j\| . \tag{4.16}$$

The energy score has its limitations, as the sample size plays a more important role with increasing dimensionality (Pinson, 2013) and it is very sensitive to errors in the mean and spread and less so to misspecified correlations (Pinson and Tastu, 2013). As an alternative, Scheuerer and Hamill (2015) propose the variogram score, which has better discrimination ability when it comes to the correlation structure:

$$\mathrm{VS}_p\left(F, y\right) = \sum_{i=1}^{d} \sum_{j=1}^{d} \omega_{ij} \left(|y_i - y_j|^p - \mathbb{E}_F \left|X_i - X_j\right|^p\right)^2 . \tag{4.17}$$

Here, X_i and X_j are components of the d-dimensional random vector X, y_i and y_j components of the observation vector and ω_{ij} optional non-negative weights. The order p can be chosen freely, but it is recommended to be set to $p = 0.5$.

It is again important to use multiple scores in order to have more information about the different aspects of forecast skill. As with the univariate scores, the sample size should be sufficiently large and confidence intervals should be given.

4.4 Comparing probability distributions

In each forecast setting previously mentioned in this chapter, we try to predict one observation value as accurately as possible. Climate projections, for example, operate differently: here, the focus is on matching the observed climate, which is the distribution of the observations over a long time period. This means that we no longer compare an ensemble or a distribution to a single deterministic value, but rather two distributions.

To this end, Thorarinsdottir, Gneiting, and Gissibl (2013) derive divergence functions from proper scores that are themselves proper. Examples for these are the integrated quadratic distance for a predictive distribution F, relative to the observed distribution G, which is based on the CRPS:

$$\mathrm{IQD}\left(F, G\right) = \int_{-\infty}^{\infty} \left(F\left(x\right) - G\left(x\right)\right)^2 \,\mathrm{d}x. \tag{4.18}$$

The score divergence associated with the ignorance score, the Kullback-Leibler divergence, can break down if the observation distribution is given as an empirical distribution, which is usually the case. Hence, this score divergence is rather impractical to use.

Some applications might require a focus on certain aspects of a distribution. A score divergence derived from the squared error, the mean value divergence, only compares the means of two distributions:

$$\mathrm{MVD}\left(F, G\right) = \left(\mathrm{mean}\left(F\right) - \mathrm{mean}\left(G\right)\right)^2 . \tag{4.19}$$

The Brier divergence is based on the Brier score and assesses the skill relative to a certain threshold u:

$$\mathrm{BD}\left(F, G \mid u\right) = \left(G\left(u\right) - F\left(u\right)\right)^2 . \tag{4.20}$$

Figure 6.12 in Paper IV shows some examples for the different score divergences and how they compare in terms of discrimination ability.

Chapter 5

Summary of papers

Paper I introduces rapid adjustment of forecast trajectories (RAFT), a new family of post-processing methods. Weather forecasts are usually not revisited once issued and are only updated when a new run of the numerical weather prediction (NWP) model is available. We propose to use error information from the part of the trajectory that has already verified to enhance forecast skill at future lead times. To this end, we utilise the error correlation between forecast lead times and link them with a linear regression model. The regression coefficients allow us to define an adjustment period, i.e., the time period preceding each forecast lead time where errors are correlated enough so that an adjustment will result in a genuine improvement in forecast skill.

RAFT is applied to temperature forecasts from the UK Met Office's MOGREPS-UK convective-scale ensemble (Hagelin et al., 2017), which have been previously post-processed with the state-of-art ensemble model output statistics (EMOS) technique (Gneiting et al., 2005). In this way, large deterministic and probabilistic biases are removed, before increasing the skill of the ensemble mean further with RAFT. Both methods complement each other, as they pertain to different stages of the post-processing chain. While EMOS is carried out once, as soon as the NWP model has finished its run, we apply RAFT adjustments at every forecast time step to the remaining part of the trajectory.

Results are shown for Heathrow Airport and 149 other locations in the UK and the Republic of Ireland. By adding the latest information about the future prediction error, forecast trajectories become substantially more accurate, with the most gain in skill realised in the final adjustment steps. Although predictability of surface temperature forecasts varies greatly with the diurnal cycle, RAFT compensates for periods of low predictive skill during nighttime. Adjusted older model runs become so much more skilful that they outperform the newest run for a brief time.

RAFT deterministic forecasts may be used as an updated mean in the EMOS predictive distribution. Although the sign of miscalibration changes from slightly under- to slightly overdispersed, the new forecast distributions are closer to perfect calibration. Therefore, the combined distributions of RAFT mean and EMOS variance provide an improved probabilistic forecast, in addition to the reduced deterministic bias.

Paper II shows how RAFT can be applied to individual ensemble members in order to adjust the ensemble mean and variance simultaneously. MOGREPS-UK surface wind speed forecasts are post-processed with two different EMOS variants, using truncated Gaussian (Thorarinsdottir and Gneiting, 2010) and truncated logistic distributions (Scheuerer and Möller, 2015), respectively. Both techniques perform similarly well, with one being slightly better calibrated and the other producing more accurate mean forecasts.

Typically, statistical post-processing addresses locations, weather parameters and forecast lead times separately. As a consequence, physical dependencies between these variables are destroyed and forecasts can become inconsistent when viewed in a multivariate sense. Techniques like ensemble copula coupling (ECC; Schefzik, Thorarinsdottir, and Gneiting, 2013) copy the raw ensemble's correlation structure and transfer it to the post-processed forecasts. In this paper, we try to find an answer to the question in which order the three different techniques, EMOS, RAFT and ECC, should be combined in order to build a comprehensive multi-stage post-processing system.

We investigate two alternatives: EMOS + $\mathrm{RAFT_m}$ + ECC, where ECC is applied to ensemble members sampled from the EMOS predictive distribution with RAFT-adjusted mean, and EMOS + ECC + $\mathrm{RAFT_{ens}}$, where we sample from the original EMOS distribution, apply ECC and then adjust the ensemble members with the new RAFT version. Our findings show that the two options for combinations of post-processing methods receive similar overall scores, but one might decide to implement one version and not the other, depending on the target statistics.

If focussing on univariate verification tools, EMOS + $\mathrm{RAFT_m}$ + ECC is the better choice, as it performs best in terms of calibration and the continuous ranked probability score (CRPS), while the root-mean-square error (RMSE) values are roughly equal. As with temperature, the MOGREPS-UK wind speed forecasts are still underdispersed after the EMOS step, which is compensated by applying RAFT$_\mathrm{m}$. For multivariate coherency, however, EMOS + ECC + RAFT$_\mathrm{ens}$ should be preferred, as it receives lower energy scores and Euclidean errors. This post-processing chain is also less dependent on the level of calibration after completing the initial EMOS step. If forecasts are perfectly calibrated or overdispersed, the RAFT$_\mathrm{ens}$ version will most likely also produce the more skilful univariate forecasts.

Paper III illustrates that RAFT can work on multiple NWP time scales, as long as the forecast lead times are correlated. It is also demonstrated how RAFT complements lagged ensembles, a collection of ensemble members from different model runs, by updating the older members to the skill level of the newest ones. A particular challenge in weather forecasting are sub-seasonal to seasonal predictions, as they rely on specific low-frequency patterns to provide predictability beyond the usual two-week limit of synoptic-scale forecasting (e.g., Vitart and Robertson, 2019). At this time scale, forecasts should be of a probabilistic nature and skilful ensembles are essential. We apply RAFT to the UK Met Office's GloSea5 coupled ocean-atmosphere seasonal prediction system (MacLachlan et al., 2015) for spring and summer temperature anomalies in Europe, using the ERA5 reanalysis (Copernicus Climate Change Service, 2017) for post-processing and verification. In order to obtain a sufficiently large sample, GloSea5's hindcast data set is used, which covers 23 years from 1993 to 2015. Due to limitations in the temporal correlation structure, we concentrate on weekly average forecasts.

Some changes have to be made to the original RAFT technique that was designed for short-range weather models. The adjustment period is determined subjectively, as we have to take care to avoid spurious correlations, and we now treat the whole forecast grid simultaneously, rather than individual sites. Five model runs initialised between 1 May and 1 June are considered, producing forecasts for a maximum of 18 weeks until the beginning of September.

Overall, a seasonal forecasting system should outperform a climatological forecast in order to be useful. Thus, we assess the forecast quality in terms of the skill score of the RMSE with the ERA5 climatology as reference forecast. The individual model runs are for the first two weeks considerably more skilful than climatology, but drop to about 5 to 15% below climatology afterwards. Adjusted ensemble means from the final RAFT iteration improve the original forecasts, such that they are more skilful than climatology for later lead times.

Although the ensemble means of every single model run have lower skill than climatology, the mean of the lagged ensemble performs on roughly the same level, while the skill score of the respective RAFT combined mean is consistently better than climatology. We can also adjust the climatological forecast in the same manner as the ensemble; these forecasts are then 5 to 10% more skilful than the unadjusted climatology. For obtaining the largest skill, the RAFT climatology forecast should be included in the lagged ensemble. While the benefit of applying RAFT to lagged ensembles is shown, the correlation between lead times only allows for adjustments one or two weeks ahead, which limits the usefulness of the adjusted forecasts.

Paper IV provides an overview of verification tools commonly used for assessing the calibration and sharpness of probabilistic weather forecasts, as well as deterministic forecasts derived from ensembles. Calibration is the statistical compatibility between forecasts and observations, and sharpness refers to the concentration of the predictive distribution. The goal of any forecasting system should be to maximise the sharpness subject to calibration (Gneiting, Balabdaoui, and Raftery, 2007).

The most widely used diagnostic tools for univariate calibration are the verification rank (Anderson, 1996; Hamill and Colucci, 1997) and PIT histograms (Dawid, 1984; Diebold, Gunther, and Tay, 1998). In the multivariate case, there are several options like the multivariate rank histogram (Gneiting et al., 2008), the average rank and band depth histograms (both Thorarinsdottir, Scheuerer, and Heinz, 2016) and the minimum spanning tree histogram (Smith and Hansen, 2004; Wilks, 2004). We compare these methods and show how they can be interpreted by means of an example.

Proper scoring rules are important techniques to assess forecast accuracy; different forecasters can easily be ranked, as the scores always reward the true predictive distribution with the lowest value (Gneiting and Raftery, 2007). In a simulation study, we test if scores agree on how a set of different forecast distributions should be ranked. For this purpose, we repeat the experiment twice, once with a sample size that is realistic for most research studies, and once with 1000 times as many data points to find the true ranking. We conclude that inference about the best forecaster is difficult for the smaller data set and mean scores should always be accompanied by some quantification of uncertainty, e.g., confidence intervals. Different scores can highlight different aspects of forecast quality, thus it is advised to use a combination of scoring rules to compare forecasting methods.

Another focus in this paper lies on the verification of extremes with proper scores. The so-called “forecaster’s dilemma” describes how restricting the observations to subsets can lead to hedging strategies, even when using scores that are otherwise proper (Lerch et al., 2017). In a similar setting to the previous simulation study, it is shown how weighted versions of proper scores can be constructed and applied in order to identify the best forecaster for extreme values.

Multivariate scores also take into account the correlation between components of multidimensional forecasts; we discuss their respective strengths and weaknesses. In the context of climate modelling, it is necessary to compare two distributions, specifically the predictive distribution to the distribution of observations. For this purpose, divergence functions that are associated with proper scores have been proposed (Thorarinsdottir, Gneiting, and Gissibl, 2013). Here, we give an overview of these divergence functions and show their respective properties in an example. We conclude by commenting on tools commonly used to understand a NWP model’s performance. Although they are sometimes not proper and typically should not be used to rank competing forecasters, there is value in applying these methods to investigate certain aspects of a forecasting system.

Chapter 6

Conclusions

In this thesis, we have shown how forecast trajectories can be improved even after they have been issued. As soon as observations for the first few time steps are recorded, we adjust the remaining part of the trajectory based on new error information. The empirical correlation structure of forecast errors allows us to connect different lead times and define an adjustment period, during which an adjustment results in increased forecast skill. Even when applied on top of state-of-the-art statistical post-processing like ensemble model output statistics (EMOS), rapid adjustment of forecast trajectories (RAFT) improves on the already calibrated forecasts by adding information about the systematic error that was previously not available.

Studies for surface temperature and wind speed forecasts from the convection-permitting MOGREPS-UK ensemble demonstrate the benefit of two different RAFT approaches, one where only the mean forecast is adjusted and another where all ensemble members, and thus the ensemble spread, are updated. For this specific ensemble, the mean forecast version performs slightly better, as the original forecasts were underdispersed. Although not adjusting the variance results in overdispersion, the combined RAFT/EMOS distributions are nonetheless closer to being perfectly calibrated. On the other hand, the accuracy of the deterministic forecast is optimised when using RAFT to correct individual ensemble members. The RAFT_{ens} version manages to reduce the uncertainty as new information about the development of the weather is incorporated. Caution must be taken that this does not result in underdispersed forecasts, especially for the tail of the trajectory. Other methods that rely on the same principle of increasing sharpness are shown in Raynaud and Bouttier (2015) in the form of small-scale initial perturbations for convective-scale models created with minimal computational effort, and in Dobrynin et al. (2018), who select subsamples of seasonal NAO forecast ensembles that resemble a first guess based on statistical analysis.

Many applications require forecasts to be consistent across multiple dimensions, e.g., across locations, time steps or weather parameters. Methods like ensemble copula coupling (ECC) restore the raw ensemble's correlation structure to the post-processed forecast by reordering the ensemble members. We investigated how to best combine EMOS, ECC and RAFT, each one being an integral part of a comprehensive post-processing scheme. While the results were similar overall, one might prefer the EMOS + RAFT_m + ECC combination if mainly interested in univariate performance and EMOS + ECC + RAFT_ens for a multivariate focus. The former is usually computationally more expensive, as ECC has to be applied at every update, but this varies with the number of ensemble members and forecast lead times. The computational cost of post-processing is, however, negligible in comparison to the NWP ensemble.

Whereas the application to MOGREPS-UK was not able to identify substantial differences in performance between RAFT_m and RAFT_ens, it is possible to further investigate using a synthetic data set. By letting specific characteristics of the predictive distribution vary while keeping the others fixed, the relative strengths and weaknesses of different methods can be analysed, which is especially important when multivariate forecasts are considered. Previous studies making use of such data sets include Lerch et al. (2020), who compare multivariate post-processing techniques (including the combination of EMOS and ECC), and Ben Bouallègue et al. (2016), who illustrate the effect of their modification to the ECC method.

One of the most interesting results is that RAFT-adjusted forecasts from older model runs for a brief time outperform those from the newest run. This effect is particularly important, as we have not considered the NWP model's computation time – typically several hours – in Paper I and Paper II. In a real operational setting, the time period where older, updated forecasts should be used is therefore considerably longer than presented here. Potential applications for RAFT include settings where forecasts have to be issued at a certain point in time (such as the day-ahead energy market), but the last run of the NWP model is several hours old. With RAFT, these forecasts can be updated and raised to the skill level of a recently initialised model.

In addition to the improved predictive skill, RAFT mean forecasts are also less jumpy than the EMOS and – in most cases – the raw ensemble forecasts, meaning that the forecast error reduces with successive updates while not changing sign. So far, tools analysing jumpiness or consistency mostly apply to deterministic forecasts, as it is difficult to transfer this aspect of forecast quality to predictive distributions. Richardson, Cloke, and Pappenberger (2020) recently proposed to evaluate the consistency of ensemble forecasts on the basis of the integrated quadratic distance (see Section 4.4) at different lead times. This approach does not take into account forecast accuracy and its ability to rank the consistency of competing forecasting methods is therefore quite limited. However, it points towards future research and potential ways of developing the tools required for assessing the jumpiness of forecast distributions.

The MOGREPS-UK ensemble has undergone substantial changes in recent years, among others a rapid update cycle was introduced. Instead of being initialised every six hours, a new model run is now started every hour with a reduced number of ensemble members (Met Office, 2019). An 18-member lagged ensemble, consisting of the forecasts from the last 6 cycles, is formed. These members correspond to lead times that are up to 6 hours apart and can therefore exhibit quite stark skill differences. RAFT can be used to update older members of the lagged ensemble and thus balance these discrepancies.

We examine this feature for seasonal temperature predictions of the GloSea5 long-range ensemble. There is a considerable benefit to applying RAFT in such a context, however forecasts can reasonably only be adjusted one to three weeks ahead and only for weekly aggregates, as useful correlations between lead times do not extend beyond one month. In Paper III, we only considered spring and summer temperatures and these findings might differ substantially depending on the season and weather parameter. Also, it might be more effective to apply RAFT to forecasts of atmospheric circulation patterns instead of surface weather variables.

In a way, RAFT can be regarded as in the same spirit as other methods that combine numerical weather forecasts and recent observations, such as nowcasting or short-range forecast blending (Vannitsem et al., 2020). The latter involves the estimation of blending weights, which can be achieved based on a range of criteria and with a variety of different methods (e.g., Atencia et al., 2020; Bouttier and Marchal, 2020; Schaumann et al., 2020). However, RAFT is not a direct competition of these techniques, but rather complements them. Different blending sources are often not on equal skill levels, for example a nowcast and a forecast from a NWP model run that was initialised several hours ago. With RAFT, the older sources can be updated so that they contain nearly the same amount of information as the newest predictions. In this way, computing blending weights is much easier and can result in more skilful and consistent forecasts. RAFT also has an advantage compared to post-processing techniques that use observations as persistence predictors (e.g., Hess, 2020). These methods are restricted to the data that is available when the forecast is created and persistence forecasts are usually only valid for a few hours. RAFT allows for continuous updates of the same trajectory incorporating new information – in our study up to

18 hours ahead – without creating a fully new forecast each time and using only one predictor variable, therefore reducing the computational overhead.

There are many aspects of RAFT that can still be optimised, such as the algorithm to determine the adjustment period and finding solutions for sites where local effects play a particular important role. In conclusion, we hope that RAFT will be adopted by the weather forecasting community as a straightforward, versatile and economical way to add value to forecasts even after they were issued.

References

Anderson, J. L. (1996). “A method for producing and evaluating probabilistic forecasts from ensemble model integrations”. In: *Journal of Climate* vol. 9, no. 7, pp. 1518–1530. DOI: 10.1175/1520-0442(1996)009<1518:amfpae>2.0.co;2.

Atencia, A., Wang, Y., Kann, A., and Meier, F. (2020). “Localization and flow-dependency on blending techniques”. In: *Meteorologische Zeitschrift* vol. 29, no. 3, pp. 231–246. DOI: 10.1127/metz/2019/0987.

Baran, S. (2014). “Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components”. In: *Computational Statistics & Data Analysis* vol. 75, pp. 227–238. DOI: 10.1016/j.csda.2014.02.013.

Baran, S., Horányi, A., and Nemoda, D. (2013). “Comparison of BMA and EMOS statistical calibration methods for temperature and wind speed ensemble weather prediction”. arXiv:1312.3763.

Baran, S. and Lerch, S. (2015). “Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 141, no. 691, pp. 2289–2299. DOI: 10.1002/qj.2521.

— (2016). “Mixture EMOS model for calibrating ensemble forecasts of wind speed”. In: *Environmetrics* vol. 27, no. 2, pp. 116–130. DOI: 10.1002/env.2380.

Baran, S. and Möller, A. (2014). “Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging”. In: *Environmetrics* vol. 26, no. 2, pp. 120–132. DOI: 10.1002/env.2316.

— (2016). “Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature”. In: *Meteorology and Atmospheric Physics* vol. 129, no. 1, pp. 99–112. DOI: 10.1007/s00703-016-0467-8.

Baran, S. and Nemoda, D. (2016). “Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting”. In: *Environmetrics* vol. 27, no. 5, pp. 280–292. DOI: 10.1002/env.2391.

Barnes, C., Brierley, C. M., and Chandler, R. E. (2019). “New approaches to post-processing of multi-model ensemble forecasts”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 145, no. 725, pp. 3479–3498. DOI: 10.1002/qj.3632.

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". In: *Monthly Weather Review* vol. 144, no. 12, pp. 4737–4750. DOI: 10.1175/mwr-d-15-0403.1.

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2007). "Combining spatial statistical and ensemble information in probabilistic weather forecasts". In: *Monthly Weather Review* vol. 135, no. 4, pp. 1386–1402. DOI: 10.1175/mwr3341.1.

— (2008). "Probabilistic quantitative precipitation field forecasting using a two-stage spatial model". In: *Annals of Applied Statistics* vol. 2, no. 4, pp. 1170–1193. DOI: 10.1214/08-AOAS203.

Bjerknes, V. (1904). "Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik (The problem of weather prediction, considered from the viewpoints of mechanics and physics)". In: *Meteorologische Zeitschrift* vol. 21, pp. 1–7. DOI: 10.1127/0941-2948/2009/416. (translated and edited by Volken, E. and Brönnimann, S. - Meteorologische Zeitschrift 18 (2009), 663-667).

Bouttier, F. and Marchal, H. (2020). "Probabilistic thunderstorm forecasting by blending multiple ensembles". In: *Tellus A: Dynamic Meteorology and Oceanography* vol. 72, no. 1, pp. 1–19. DOI: 10.1080/16000870.2019.1696142.

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* vol. 78, no. 1, pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

Bröcker, J. (2012). "Evaluating raw ensembles with the continuous ranked probabilty score". In: *Quarterly Journal of the Royal Meteorological Society* vol. 138, no. 667, pp. 1611–1617. DOI: 10.1002/qj.1891.

Buizza, R., Miller, M., and Palmer, T. N. (1999). "Stochastic representation of model uncertainties in the ECMWF ensemble prediction system". In: *Quarterly Journal of the Royal Meteorological Society* vol. 125, no. 560, pp. 2887–2908. DOI: 10.1002/qj.49712556006.

Buizza, R. and Palmer, T. N. (1995). "The singular vector structure of the atmospheric general circulation". In: *Journal of the Atmospheric Sciences* vol. 126, no. 9, pp. 2503–2518. DOI: 10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2.

Buizza, R. and Richardson, D. (2017). "25 years of ensemble forecasting at ECMWF". In: *ECMWF newsletter 153*. DOI: 10.21957/BV418O.

Buizza, R., Tribbia, J., Molteni, F., and Palmer, T. N. (1993). “Computation of optimal unstable structures for a numerical weather prediction model”. In: *Tellus* vol. 45A, no. 5, pp. 388–407. DOI: 10.1034/j.1600-0870.1993.t01-4-00005.x.

Charney, J. G., Fjørtoft, R., and von Neumann, J. (1950). “Numerical integration of the barotropic vorticity equation”. In: *Tellus* vol. 2, no. 4, pp. 237–254. DOI: https://doi.org/10.1111/j.2153-3490.1950.tb00336.x.

Clark, M. P., Gangopadhyay, S., Hay, L. E., Rajagopalan, B., and Wilby, R. L. (2004). “The Schaake Shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields”. In: *Journal of Hydrometeorology* vol. 5, pp. 243–262. DOI: 10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Copernicus Climate Change Service (2017). *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate.* URL: https://cds.climate.copernicus.eu/cdsapp#!/home.

— (2020). *Seasonal forecasts.* last accessed 01-05-2020. URL: https://climate.copernicus.eu/seasonal-forecasts.

Dawid, A. P. (1984). “Present position and potential developments: Some personal views: Statistical theory: The prequential approach”. In: *Journal of the Royal Statistical Society Series A (General)* vol. 147, no. 2, pp. 278–292. DOI: 10.2307/2981683.

Demaeyer, J. and Vannitsem, S. (2020). “Correcting for model changes in statistical post-processing - an approach based on response theory”. In: *Nonlinear Processes in Geophysics* vol. 27, no. 2, pp. 307–327. DOI: 10.5194/npg-27-307-2020.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* vol. 39B, pp. 1–39. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). “Evaluating density forecasts with applications to financial risk management”. In: *International Economic Review* vol. 39, no. 4, pp. 863–883. DOI: 10.2307/2527342.

Diebold, F. X. and Mariano, R. S. (1995). “Comparing predictive accuracy”. In: *Journal of Business & Economic Statistics* vol. 13, no. 3, pp. 253–263. DOI: 10.1080/07350015.1995.10524599.

Diks, C., Panchenko, V., and van Dijk, D. (2011). “Likelihood-based scoring rules for comparing density forecasts in tails”. In: *Journal of Econometrics* vol. 163, no. 2, pp. 215–230. DOI: 10.1016/j.jeconom.2011.04.001.

Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L. R. L., and van Oldenborgh, G. J. (2013a). “Initialized near-term regional climate change prediction”. In: *Nature Communications* vol. 4, p. 1715. DOI: 10.1038/ncomms2704.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. L. (2013b). “Seasonal climate predictability and forecasting: Status and prospects”. In: *Wiley Interdisciplinary Reviews: Climate Change* vol. 4, no. 4, pp. 245–268. DOI: 10.1002/wcc.217.

Dobrynin, M., Domeisen, D. I. V., Müller, W. A., Bell, L., Brune, S., Bunzel, F., Düsterhus, A., Fröhlich, K., Pohlmann, H., and Baehr, J. (2018). “Improved teleconnection-based dynamical seasonal predictions of boreal winter”. In: *Geophysical Research Letters* vol. 45, no. 8, pp. 3605–3614. DOI: 10.1002/2018gl077209.

Ehret, U. (2010). “Convergence Index: A new performance measure for the temporal stability of operational rainfall forecasts”. In: *Meteorologische Zeitschrift* vol. 19, no. 5, pp. 441–451. DOI: 10.1127/0941-2948/2010/0480.

Epstein, E. (1969). “Stochastic dynamic prediction”. In: *Tellus* vol. 21, no. 6, pp. 739–759. DOI: 10.3402/tellusa.v21i6.10143.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2015). “Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression”. In: *Monthly Weather Review* vol. 143, no. 3, pp. 955–971. DOI: 10.1175/mwr-d-14-00210.1.

Feldstein, S. B. and Franzke, C. L. E. (2017). “Atmospheric teleconnection patterns”. In: *Nonlinear and Stochastic Climate Dynamics*. Ed. by Franzke, C. L. E. and O’Kane, T. J. Cambridge University Press. Chap. 3, pp. 54–104. DOI: 10.1017/9781316339251.004.

Fisher, R. A. (1922). “On the mathematical foundations of theoretical statistics”. In: *Philosophical Transactions of the Royal Society of London* vol. 222A, pp. 309–368. DOI: 10.1098/rsta.1922.0009.

Friederichs, P. and Hense, A. (2007). “Statistical downscaling of extreme precipitation events using censored quantile regression”. In: *Monthly Weather Review* vol. 135, no. 6, pp. 2365–2378. DOI: 10.1175/mwr3403.1.

Friederichs, P., Wahl, S., and Buschow, S. (2018). “Postprocessing for extreme events”. In: *Statistical postprocessing of ensemble forecasts*. Ed. by Vannitsem, S., Wilks, D. S., and Messner, J. W. Elsevier. Chap. 5, pp. 127–154. DOI: 10.1016/B978-0-12-812372-0.00005-4.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. (2018). "Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood". In: *Monthly Weather Review* vol. 146, no. 12, pp. 4323–4338. DOI: 10.1175/mwr-d-17-0364.1.

Gel, Y., Raftery, A. E., and Gneiting, T. (2004). "Calibrated probabilistic mesoscale weather field forecasting". In: *Journal of the American Statistical Association* vol. 99, no. 467, pp. 575–583. DOI: 10.1198/016214504000000872.

Glahn, H. R. and Lowry, D. A. (1972). "The use of model output statistics (MOS) in objective weather forecasting". In: *Journal of Applied Meteorology* vol. 11, no. 8, pp. 1203–1211. DOI: 10.1175/1520-0450(1972)011<1203:tuomos>2.0.co;2.

Gneiting, T. (2011). "Making and evaluating point forecasts". In: *Journal of the American Statistical Association* vol. 106, no. 494, pp. 746–762. DOI: 10.1198/jasa.2011.r10138.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). "Probabilistic forecasts, calibration and sharpness". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* vol. 69, no. 2, pp. 243–268. DOI: 10.1111/j.1467-9868.2007.00587.x.

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American Statistical Association* vol. 102, no. 477, pp. 359–378. DOI: 10.1198/016214506000001437.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation". In: *Monthly Weather Review* vol. 133, no. 5, pp. 1098–1118. DOI: 10.1175/mwr2904.1.

Gneiting, T. and Ranjan, R. (2011). "Comparing density forecasts using threshold- and quantile-weighted scoring rules". In: *Journal of Business & Economic Statistics* vol. 29, no. 3, pp. 411–422. DOI: 10.1198/jbes.2010.08110.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds". In: *TEST* vol. 17, no. 2, pp. 211–235. DOI: 10.1007/s11749-008-0114-x.

Good, I. J. (1952). "Rational decisions". In: *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 14, no. 1, pp. 107–114. DOI: 10.1111/j.2517-6161.1952.tb00104.x.

Griffiths, D., Foley, M., Ioannou, I., and Leeuwenburg, T. (2018). “Flip-flop index: Quantifying revision stability for fixed-event forecasts”. In: *Meteorological Applications* vol. 26, no. 1, pp. 30–35. DOI: 10.1002/met.1732.

Grimit, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A. (2006). “The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 132, no. 621C, pp. 2925–2942. DOI: 10.1256/qj.05.235.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N., and Tennant, W. (2017). “The Met Office convective-scale ensemble, MOGREPS-UK”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 143, no. 708, pp. 2846–2861. DOI: 10.1002/qj.3135.

Hamill, T. M. (2001). “Interpretation of rank histograms for verifying ensemble forecasts”. In: *Monthly Weather Review* vol. 129, no. 3, pp. 550–560. DOI: 10.1175/1520-0493(2001)129<0550:iorhfv>2.0.co;2.

— (2006). “Ensemble-based data assimilation”. In: *Predictability of Weather and Climate*. Ed. by Palmer, T. and Hagedorn, R. Cambridge Press. Chap. 6, pp. 124–156. DOI: 10.1017/CBO9780511617652.007.

— (2018). “Practical aspects of statistical postprocessing”. In: *Statistical postprocessing of ensemble forecasts*. Ed. by Vannitsem, S., Wilks, D. S., and Messner, J. W. Elsevier. Chap. 7, pp. 187–217. DOI: 10.1016/B978-0-12-812372-0.00007-8.

Hamill, T. M. and Colucci, S. J. (1997). “Verification of Eta-RSM short-range ensemble forecasts”. In: *Monthly Weather Review* vol. 125, no. 6, pp. 1312–1327. DOI: 10.1175/1520-0493(1997)125<1312:voersr>2.0.co;2.

Haupt, S. E., Cowie, J., Linden, S., McCandless, T., Kosovic, B., and Alessandrini, S. (2018). “Machine learning for applied weather prediction”. In: *2018 IEEE 14th International Conference on e-Science (e-Science)*. IEEE. DOI: 10.1109/escience.2018.00047.

Heinrich, C., Hellton, K. H., Lenkoski, A., and Thorarinsdottir, T. L. (2020). “Multivariate postprocessing methods for high-dimensional seasonal weather forecasts”. In: *Journal of the American Statistical Association*. In press, pp. 1–28. DOI: 10.1080/01621459.2020.1769634.

Hemri, S., Lisniak, D., and Klein, B. (2015). “Multivariate postprocessing techniques for probabilistic hydrological forecasting”. In: *Water Resources Research* vol. 51, no. 9, pp. 7436–7451. DOI: 10.1002/2014wr016473.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and Haiden, T. (2014). “Trends in the predictive performance of raw ensemble weather forecasts”. In: *Geophysical Research Letters* vol. 41, no. 24, pp. 9197–9205. DOI: 10.1002/2014GL062472.

Hersbach, H. (2000). “Decomposition of the continuous ranked probability score for ensemble prediction systems”. In: *Weather and Forecasting* vol. 15, no. 5, pp. 559–570. DOI: 10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2.

Hess, R. (2020). “Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst”. In: *Nonlinear Processes in Geophysics Discussion*. In review. DOI: 10.5194/npg-2019-64.

Hoskins, B. (2012). “The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 139, no. 672, pp. 573–584. DOI: 10.1002/qj.1991.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression.* John Wiley & Sons, Inc. ISBN: 9780470582473. DOI: 10.1002/9781118548387.

Hu, Y., Schmeits, M. J., van Andel, S. J., 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”. In: *Journal of Hydrometeorology* vol. 17, no. 9, pp. 2405–2417. DOI: 10.1175/jhm-d-15-0205.1.

Joslyn, S. L. and LeClerc, J. E. (2012). “Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error.” In: *Journal of Experimental Psychology: Applied* vol. 18, no. 1, pp. 126–140. DOI: 10.1037/a0025185.

Kalnay, E. (2002). *Atmospheric modeling, data assimilation and predictability.* Cambridge University Press. ISBN: 9780511802270. DOI: 10.1017/CBO9780511802270.

Klemm, T. and McPherson, R. A. (2017). “The development of seasonal climate forecasting for agricultural producers”. In: *Agricultural and Forest Meteorology* vol. 232, pp. 384–399. DOI: 10.1016/j.agrformet.2016.09.005.

Lahiri, S. N. (2003). *Resampling methods for dependent data.* Springer. ISBN: 9781475738032. DOI: 10.1007/978-1-4757-3803-2.

Laio, F. and Tamea, S. (2007). “Verification tools for probabilistic forecasts of continuous hydrological variables”. In: *Hydrology and Earth System Sciences* vol. 11, no. 4, pp. 1267–1277. DOI: 10.5194/hess-11-1267-2007.

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”. In: *Nonlinear Processes in Geophysics* vol. 27, no. 1, pp. 23–34. DOI: 10.5194/npg-27-23-2020.

Leith, C. E. (1974). “Theoretical skill of Monte Carlo forecasts”. In: *Monthly Weather Review* vol. 102, no. 6, pp. 409–418. DOI: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

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”. In: *Nonlinear Processes in Geophysics* vol. 27, no. 2, pp. 349–371. DOI: 10.5194/npg-27-349-2020.

Lerch, S. and Thorarinsdottir, T. L. (2013). “Comparison of non-homogeneous regression models for probabilistic wind speed forecasting”. In: *Tellus A: Dynamic Meteorology and Oceanography* vol. 65, no. 1, p. 21206. DOI: 10.3402/tellusa.v65i0.21206.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). “Forecaster’s dilemma: Extreme events and forecast evaluation”. In: *Statistical Science* vol. 32, no. 1, pp. 106–127. DOI: 10.1214/16-sts588.

Lewis, J. M. (2005). “Roots of ensemble forecasting”. In: *Monthly Weather Review* vol. 133, no. 7, pp. 1865–1885. DOI: 10.1175/MWR2949.1.

— (2014). “Edward Epstein’s stochastic-dynamic approach to ensemble weather prediction”. In: *Bulletin of the American Meteorological Society* vol. 95, no. 1, pp. 99–116. DOI: 10.1175/BAMS-D-13-00036.1.

Lledó, L., Cionni, I., Torralba, V., Bretonnière, P.-A., and Samsó, M. (2020). “Seasonal prediction of Euro-Atlantic teleconnections from multiple systems”. In: *Environmental Research Letters* vol. 15, no. 7, p. 074009. DOI: 10.1088/1748-9326/ab87d2.

Lorenz, E. N. (1963). “Deterministic nonperiodic flow”. In: *Journal of the Atmospheric Sciences* vol. 20, no. 2, pp. 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

— (1969). “The predictability of a flow which possesses many scales of motions”. In: *Tellus* vol. 21, no. 3, pp. 289–307. DOI: 10.1111/j.2153-3490.1969.tb00444.x.

Lynch, P. (2006). *The emergence of numerical weather prediction: Richardson’s dream.* Cambridge University Press. ISBN: 9780521857291.

— (2008). “The ENIAC forecasts: A re-creation”. In: *Bulletin of the American Meteorological Society* vol. 89, no. 1, pp. 45–56. DOI: 10.1175/BAMS-89-1-45.

MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A. A., Gordon, M., Vellinga, M., Williams, A., Comer, R. E., Camp, J., Xavier, P., and Madec, G. (2015). “Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 141, no. 689, pp. 1072–1084. DOI: 10.1002/qj.2396.

MacLeod, D. A., Jones, A., Di Giuseppe, F., Caminade, C., and Morse, A. P. (2015). "Demonstration of successful malaria forecasts for Botswana using an operational seasonal climate model". In: *Environmental Research Letters* vol. 10, no. 4, p. 044005. DOI: 10.1088/1748-9326/10/4/044005.

Matheson, J. E. and Winkler, R. L. (1976). "Scoring rules for continuous probability distributions". In: *Management Science* vol. 22, no. 10, pp. 1087–1096. DOI: 10.1287/mnsc.22.10.1087.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions.* John Wiley & Sons, Inc. ISBN: 9780470191613. DOI: 10.1002/9780470191613.

Met Office (2019). *Improvements to the UK ensemble.* last accessed 20-01-2020. URL: https://www.metoffice.gov.uk/research/news/2019/mogreps-uk-hourly-cycling-updates.

— (2020). *Unified Model.* last accessed 20-04-2020. URL: https://www.metoffice.gov.uk/research/approach/modelling-systems/unified-model/index.

Metropolis, N. and Ulam, S. (1949). "The Monte Carlo Method". In: *Journal of the American Statistical Association* vol. 44, no. 247, pp. 335–341. DOI: 10.1080/01621459.1949.10483310.

Möller, A., Lenkoski, A., and Thorarinsdottir, T. L. (2013). "Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas". In: *Quarterly Journal of the Royal Meteorological Society* vol. 139, no. 673, pp. 982–991. DOI: 10.1002/qj.2009.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). "The ECMWF ensemble prediction system: methodology and validation". In: *Quarterly Journal of the Royal Meteorological Society* vol. 122, no. 529, pp. 73–119. DOI: 10.1002/qj.49712252905.

Murphy, A. H. (1973). "Hedging and skill scores for probability forecasts". In: *Journal of Applied Meteorology* vol. 12, no. 1, pp. 215–223. DOI: 10.1175/1520-0450(1973)012<0215:hassfp>2.0.co;2.

— (1993). "What is a good forecast? An essay on the nature of goodness in weather forecasting". In: *Weather and Forecasting* vol. 8, no. 2, pp. 281–293. DOI: 10.1175/1520-0434(1993)008<0281:wiagfa>2.0.co;2.

Novak, D. R., Bailey, C., Brill, K., Eckert, M., Petersen, D., Rausch, R., and Schichtel, M. (2011). "Human improvement to numerical weather prediction at the Hydrometeorological Prediction Center". In: *Proceedings of the 24th Conference on Weather Analysis and Forecasting/20th Conference on Numerical Weather Prediction.*

Orlov, A., Sillmann, J., and Vigo, I. (2020). “Better seasonal forecasts for the renewable energy industry”. In: *Nature Energy* vol. 5, no. 2, pp. 108–110. DOI: 10.1038/s41560-020-0561-5.

Pinson, P. (2012). “Adaptive calibration of (u,v)-wind ensemble forecasts”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 138, no. 666, pp. 1273–1284. DOI: 10.1002/qj.1873.

— (2013). “Wind energy: Forecasting challenges for its operational management”. In: *Statistical Science* vol. 28, no. 4, pp. 564–585. DOI: 10.1214/13-sts445.

Pinson, P. and Girard, R. (2012). “Evaluating the quality of scenarios of short-term wind power generation”. In: *Applied Energy* vol. 96, pp. 12–20. DOI: 10.1016/j.apenergy.2011.11.004.

Pinson, P. and Tastu, J. (2013). *Discrimination ability of the energy score.* Tech. rep. Technical University of Denmark (DTU).

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). “Using Bayesian model averaging to calibrate forecast ensembles”. In: *Monthly Weather Review* vol. 133, no. 5, pp. 1155–1174. DOI: 10.1175/MWR2906.1.

Raynaud, L. and Bouttier, F. (2015). “Comparison of initial perturbation methods for ensemble prediction at convective scale”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 142, no. 695, pp. 854–866. DOI: 10.1002/qj.2686.

Richardson, D. S., Cloke, H. L., and Pappenberger, F. (2020). “Evaluation of the consistency of ECMWF ensemble forecasts”. In: *Geophysical Research Letters* vol. 46, e2020GL087934. DOI: 10.1029/2020gl087934.

Richardson, L. F. (1922). *Weather prediction by numerical process.* Cambridge University Press. ISBN: 9780511618291.

Robertson, A. W. and Vitart, F. (2019). “Epilogue”. In: *Sub-Seasonal to Seasonal Prediction.* Ed. by Robertson, A. W. and Vitart, F. Elsevier. Chap. 23, pp. 479–481. DOI: 10.1016/b978-0-12-811714-9.00023-1.

Ruth, D. P., Glahn, B., Dagostaro, V., and Gilbert, K. (2009). “The performance of MOS in the digital age”. In: *Weather and Forecasting* vol. 24, no. 2, pp. 504–519. DOI: 10.1175/2008waf2222158.1.

Schaumann, P., de Langlard, M., Hess, R., James, P., and Schmidt, V. (2020). “A calibrated combination of probabilistic precipitation forecasts to achieve a seamless transition from nowcasting to very short-range forecasting”. In: *Weather and Forecasting* vol. 35, no. 3, pp. 773–791. DOI: 10.1175/waf-d-19-0181.1.

Schefzik, R. (2016). “A similarity-based implementation of the Schaake Shuffle”. In: *Monthly Weather Review* vol. 144, no. 5, pp. 1909–1921. DOI: 10.1175/mwr-d-15-0227.1.

— (2017). “Ensemble calibration with preserved correlations: unifying and comparing ensemble copula coupling and member-by-member postprocessing”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 143, no. 703, pp. 999–1008. DOI: 10.1002/qj.2984.

Schefzik, R. and Möller, A. (2018). “Ensemble postprocessing methods incorporating dependence structures”. In: *Statistical postprocessing of ensemble forecasts*. Ed. by Vannitsem, S., Wilks, D. S., and Messner, J. W. Elsevier. Chap. 4, pp. 91–125. DOI: 10.1016/B978-0-12-812372-0.00004-2.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. (2013). “Uncertainty quantification in complex simulation models using ensemble copula coupling”. In: *Statistical Science* vol. 28, no. 4, pp. 616–640. DOI: 10.1214/13-STS443.

Scheuerer, M. (2013). “Probabilistic quantitative precipitation forecasting using ensemble model output statistics”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 140, no. 680, pp. 1086–1096. DOI: 10.1002/qj.2183.

Scheuerer, M. and Büermann, L. (2014). “Spatially adaptive post-processing of ensemble forecasts for temperature”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* vol. 63, no. 3, pp. 405–422. DOI: 10.1111/rssc.12040.

Scheuerer, M. and Hamill, T. M. (2015). “Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities”. In: *Monthly Weather Review* vol. 143, no. 4, pp. 1321–1334. DOI: 10.1175/mwr-d-14-00269.1.

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”. In: *Water Resources Research* vol. 53, no. 4, pp. 3029–3046. DOI: 10.1002/2016wr020133.

Scheuerer, M. and Möller, D. (2015). “Probabilistic wind speed forecasting on a grid based on ensemble model output statistics”. In: *Annals of Applied Statistics* vol. 9, no. 3, pp. 1328–1349. DOI: 10.1214/15-aoas843.

Schmeits, M. J. and Kok, K. J. (2010). “A comparison between raw ensemble output, (Modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts”. In: *Monthly Weather Review* vol. 138, no. 11, pp. 4199–4211. DOI: 10.1175/2010mwr3285.1.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periañez, A., and Potthast, R. (2016). "Kilometre-scale ensemble data assimilation for the COSMO model (KENDA)". In: *Quarterly Journal of the Royal Meteorological Society* vol. 142, no. 696, pp. 1453–1472. DOI: 10.1002/qj.2748.

Schuhen, N., Buchanan, P., Evans, G., and Jackson, S. (2016). "A comparative study of statistical post-processing methods for the calibration of ensemble forecasts". In: *96th Annual Meeting of the American Meteorological Society*. New Orleans, LA, USA.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T. (2012). "Ensemble model output statistics for wind vectors". In: *Monthly Weather Review* vol. 140, no. 10, pp. 3204–3219. DOI: 10.1175/mwr-d-12-00028.1.

Sharpe, M. A., Bysouth, C. E., and Stretton, R. L. (2017). "How well do Met Office post-processed site-specific probabilistic forecasts predict relative-extreme events?" In: *Meteorological Applications* vol. 25, no. 1, pp. 23–32. DOI: 10.1002/met.1665.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010). "Probabilistic wind speed forecasting using ensembles and Bayesian model averaging". In: *Journal of the American Statistical Association* vol. 105, no. 489, pp. 25–35. DOI: 10.1198/jasa.2009.ap08615.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2013). "Probabilistic wind vector forecasting using ensembles and Bayesian model averaging". In: *Monthly Weather Review* vol. 141, no. 6, pp. 2107–2119. DOI: 10.1175/mwr-d-12-00002.1.

Sloughter, J. M., Raftery, A. E., Gneiting, T., and Fraley, C. (2007). "Probabilistic quantitative precipitation forecasting using Bayesian model averaging". In: *Monthly Weather Review* vol. 135, no. 9, pp. 3209–3220. DOI: 10.1175/mwr3441.1.

Smith, L. A. and Hansen, J. A. (2004). "Extending the limits of ensemble forecast verification with the minimum spanning tree". In: *Monthly Weather Review* vol. 132, no. 6, pp. 1522–1528. DOI: 10.1175/1520-0493(2004)132<1522:etloef>2.0.co;2.

Steiner, A., Köhler, C., Metzinger, I., Braun, A., Zirkelbach, M., Ernst, D., Tran, P., and Ritter, B. (2017). "Critical weather situations for renewable energies – Part A: Cyclone detection for wind power". In: *Renewable Energy* vol. 101, pp. 41–50. DOI: 10.1016/j.renene.2016.08.013.

Talagrand, O., Vautard, R., and Strauss, B. (1997). "Evaluation of probabilistic prediction systems". In: *Proc. Workshop on Predictability*. Reading, UK, European Centre for Medium-Range Weather Forecasts, pp. 1–25. URL: https://www.ecmwf.int/node/12555.

Tennant, W. J., Shutts, G. J., Arribas, A., and Thompson, S. A. (2011). "Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill". In: *Monthly Weather Review* vol. 139, no. 4, pp. 1190–1206. DOI: 10.1175/2010MWR3430.1.

Thorarinsdottir, T. L. and Gneiting, T. (2010). "Probabilistic forecasts of wind speed: Ensemble model output statistics using heteroskedastic censored regression". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* vol. 173, no. 2, pp. 371–388. DOI: 10.1111/j.1467-985X.2009.00616.x.

Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N. (2013). "Using proper divergence functions to evaluate climate models". In: *SIAM/ASA Journal on Uncertainty Quantification* vol. 1, no. 1, pp. 522–534. DOI: 10.1137/130907550.

Thorarinsdottir, T. L. and Johnson, M. S. (2012). "Probabilistic wind gust forecasting using non-homogeneous Gaussian regression". In: *Monthly Weather Review* vol. 140, no. 3, pp. 889–897. DOI: 10.1175/MWR-D-11-00075.1.

Thorarinsdottir, T. L., Scheuerer, M., and Heinz, C. (2016). "Assessing the calibration of high-dimensional ensemble forecasts using rank histograms". In: *Journal of Computational and Graphical Statistics* vol. 25, no. 1, pp. 105–122. DOI: 10.1080/10618600.2014.977447.

Toth, Z. and Kalnay, E. (1993). "Ensemble forecasting at NMC: The generation of perturbation". In: *Bulletin of the American Meteorological Society* vol. 74, no. 12, pp. 2317–2330. DOI: 10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Tracton, M. S. and Kalnay, E. (1993). "Operational ensemble prediction at the National Meteorological Center: Practical aspects". In: *Weather and Forecasting* vol. 8, no. 3, pp. 379–398. DOI: 10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

Van Schaeybroeck, B. and Vannitsem, S. (2014). "Ensemble post-processing using member-by-member approaches: theoretical aspects". In: *Quarterly Journal of the Royal Meteorological Society* vol. 141, no. 688, pp. 807–818. DOI: 10.1002/qj.2397.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Bouallègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Plenković, I. O., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K., and Ylhaisi, J. (2020). "Statistical postprocessing for weather forecasts – review, challenges and avenues in a Big Data world". arXiv:2004.06582.

Vitart, F. and Robertson, A. W. (2019). “Introduction: Why sub-seasonal to seasonal prediction (S2S)?” In: *Sub-Seasonal to Seasonal Prediction*. Ed. by Robertson, A. W. and Vitart, F. Elsevier. Chap. 1, pp. 3–15. DOI: 10.1016/b978-0-12-811714-9.00001-2.

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. A., and Palmer, T. N. (2017). “Atmospheric seasonal forecasts of the twentieth century: Multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 143, no. 703, pp. 917–926. DOI: 10.1002/qj.2976.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A. P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N. J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T. J., Street, R., Jones, L., Remenyi, T. A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., and Zebiak, S. E. (2017). “Potential applications of subseasonal-to-seasonal (S2S) predictions”. In: *Meteorological Applications* vol. 24, no. 3, pp. 315–325. DOI: 10.1002/met.1654.

Wilks, D. S. (2004). “The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts”. In: *Monthly Weather Review* vol. 132, no. 6, pp. 1329–1340. DOI: 10.1175/1520-0493(2004)132<1329:tmstha>2.0.co;2.

— (2011). *Statistical methods in the atmospheric sciences*. Elsevier/Academic Press. ISBN: 9780123850232.

— (2014). “Multivariate ensemble model output statistics using empirical copulas”. In: *Quarterly Journal of the Royal Meteorological Society* vol. 141, no. 688, pp. 945–952. DOI: 10.1002/qj.2414.

— (2018). “Univariate ensemble postprocessing”. In: *Statistical postprocessing of ensemble forecasts*. Ed. by Vannitsem, S., Wilks, D. S., and Messner, J. W. Elsevier. Chap. 3, pp. 49–89. DOI: 10.1016/B978-0-12-812372-0.00003-0.

Zhang, F., Sun, Y. Q., Magnusson, L., Buizza, R., Lin, S.-J., Chen, J.-H., and Emanuel, K. (2019). “What Is the predictability limit of midlatitude weather?” In: *Journal of the Atmospheric Sciences* vol. 76, no. 4, pp. 1077–1091. DOI: 10.1175/jas-d-18-0269.1.

Zsoter, E., Buizza, R., and Richardson, D. (2009). ““Jumpiness” of the ECMWF and Met Office EPS control and ensemble-mean forecasts”. In: *Monthly Weather Review* vol. 137, no. 11, pp. 3823–3836. DOI: 10.1175/2009mwr2960.1.

Papers

Paper I

Rapid adjustment and post-processing of temperature forecast trajectories

Nina Schuhen, Thordis L. Thorarinsdottir, Alex Lenkoski

In: Quarterly Journal of the Royal Meteorological Society,
vol. 146, no. 727, pp. 963-978, 2020.
DOI: 10.1002/qj.3718

Received: 4 October 2019 | Revised: 20 November 2019 | Accepted: 2 December 2019 | Published on: 7 January 2020

DOI: 10.1002/qj.3718

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Rapid adjustment and post-processing of temperature forecast trajectories

N. Schuhen | T. L. Thorarinsdottir | A. Lenkoski

Norwegian Computing Center, Oslo, Norway

Correspondence
N. Schuhen, Norwegian Computing Center, PO Box 114, Blindern, NO-0314 Oslo, Norway.
Email:nina.schuhen@nr.no

Funding information
Research Council of Norway, 259864, 237718 and 270733

Abstract

Modern weather forecasts are commonly issued as consistent multi-day forecast trajectories with a time resolution of 1–3 hours. Prior to issuing, statistical post-processing is routinely used to correct systematic errors and misrepresentations of the forecast uncertainty. However, once the forecast has been issued, it is rarely updated before it is replaced in the next forecast cycle of the numerical weather prediction (NWP) model. This paper shows that the error correlation structure within the forecast trajectory can be utilized to substantially improve the forecast between the NWP forecast cycles by applying additional post-processing steps each time new observations become available. The proposed rapid adjustment is applied to temperature forecast trajectories from the UK Met Office's convective-scale ensemble MOGREPS-UK. MOGREPS-UK is run four times daily and produces hourly forecasts for up to 36 hours ahead. Our results indicate that the rapidly adjusted forecast from the previous NWP forecast cycle outperforms the new forecast for the first few hours of the next cycle, or until the new forecast itself can be rapidly adjusted, suggesting a new strategy for updating the forecast cycle.

KEYWORDS
atmosphere, ensembles, forecasting (methods), statistical methods

1 | INTRODUCTION

Weather forecasts resulting from numerical weather prediction (NWP) models are traditionally post-processed using statistical approaches in order to correct potential systematic biases in the forecasts (Glahn and Lowry, 1972). Roughly 15 years ago, the first papers on statistical post-processing methods yielding full predictive distributions – correcting both systematic biases and assessments of forecast uncertainty – appeared in the literature (Gneiting *et al.,* 2005; Raftery *et al.,* 2005). Since then, approaches of this type have become increasingly more common in both the literature and operational forecasting for NWP forecasts and forecast ensembles (Vannitsem *et al.,* 2018). Originally, the methods applied to marginal predictive distributions of individual weather variables at individual locations (Gneiting *et al.,* 2005; Raftery *et al.,* 2005). More recent work has produced consistent probabilistic predictions for temporal trajectories (Hemri *et al.,* 2015), spatial forecast fields (Berrocal *et al.,* 2008; Feldmann *et al.,* 2015) and multiple variables (Schuhen *et al.,* 2012; Möller *et al.,* 2013; Sloughter *et al.,* 2013). Vannitsem *et al.,* (2018) gives a recent overview of statistical post-processing methods for ensemble forecasts.

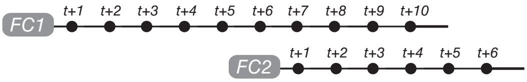


FIGURE 1 Diagram of a typical forecast cycle for hourly forecasts issued every 6 hr. The MOGREPS-UK version used in this paper is configured in this way

The aim of probabilistic forecasting is to “maximize the sharpness of the predictive distribution subject to calibration” (Gneiting *et al.,* 2007). Here, calibration, or reliability, refers to the statistical consistency between the forecast and the observation; a forecast is (probabilistically) calibrated if events predicted to have probability P are realized with the same relative frequency in the observations. A calibrated forecast should then provide as much information regarding future weather as possible; the smaller the forecast uncertainty, or the higher the sharpness of the predictive distribution, the more information regarding future weather is contained in the forecast. In practice, the NWP model outputs a forecast trajectory for multiple lead times. As soon as the model output is available, the forecasts of the entire trajectory are post-processed using the most recent available pairs of previous forecasts and verifying observations to obtain calibrated and sharp forecasts for all lead times. A new, post-processed forecast is then issued for all future time points corresponding to the lead times of the original NWP forecast. An example of such a setting is shown in Figure 1 for an hourly forecast where a new forecast is issued every 6 hr.

In the standard setting demonstrated in Figure 1, the published forecast is not updated until it is replaced in the next forecast cycle of the NWP model. However, new information in the form of new observations becomes available every hour. In the current paper, we propose an approach for Rapid Adjustment of Forecast Trajectories (RAFT), where, in addition to standard post-processing, we regularly update the forecast every time a new piece of information becomes available by utilizing the correlation of the forecast errors within an NWP forecast trajectory. The idea behind RAFT is related to that of data assimilation, for example Mitchell and Houtekamer (2000) who developed a method to account for model error in the context of an ensemble Kalman filter technique. Here, our main priority is computational efficiency to minimize the time needed for each adjustment. We thus propose an efficient adjustment approach that is adapted to each forecast cycle, hour and lead time separately. In a case-study, we apply the method to hourly temperature forecasts from the MOGREPS-UK ensemble from the UK Met Office whose schedule follows the forecast cycle shown in Figure 1.

The remainder of the paper is organized as follows. In the next Section 2, we introduce the MOGREPS-UK (Met Office Global and Regional Ensemble Prediction System) forecast ensemble and the corresponding observations, and review the classical Ensemble Model Output Statistics (EMOS) post-processing method as well as the validation metrics used in our study. We further show the skill of the post-processed EMOS forecasts. In Section 3, we introduce our proposed method for RAFT. Results at Heathrow Airport as well as those over the entire study region are presented in the following Section 4. Finally, the paper concludes with a summary and discussion in Section 5.

2 | DATA AND CONVENTIONAL POST-PROCESSING

2.1 | MOGREPS-UK

Our dataset consists of surface temperature forecasts and observations for 150 locations in the UK and the Republic of Ireland. The forecasts are provided by the UK Met Office's convective-scale ensemble MOGREPS-UK (Hagelin *et al.,* 2017), which has been running operationally since July 2012. The dataset covers a period of 30 months between January 2014 and June 2016, during which the ensemble had a horizontal resolution of 2.2 km and produced hourly forecasts for up to 36 hr. During this time, MOGREPS-UK was run four times daily, at 0300, 0900, 1500 and 2100 UTC. The initial and boundary conditions were originally provided by the global MOGREPS-G ensemble, but since March 2016 the initial conditions have been created by adding the MOGREPS-G perturbations to the analysis of the high-resolution deterministic UK variable-resolution (UKV) model, while the boundary data continue to be provided by MOGREPS-G. The ensemble consists of one control forecast and eleven perturbed members, which we treat as twelve exchangeable ensemble members.

In this study, we consider site-specific data only, interpolated by the Met Office from model grid to observation locations. During this process, forecasts are corrected for local effects and the height differences between station and model orography. The observations are extracted from SYNOP messages at the 150 locations in Figure 2 and Met Office quality controls have been applied. We separate the data into a training set (January to December 2014) with approximately 1,300 forecast trajectories for each location, or a total of 7,018,719 forecast–observation pairs, and a test set (January 2015 to June 2016) with approximately 2,096 forecast trajectories for each location, or a total of 11,320,762 forecast–observation pairs. Although there have been several operational changes to the MOGREPS-UK model during these periods, we treat the dataset as homogeneous over the entire study period.

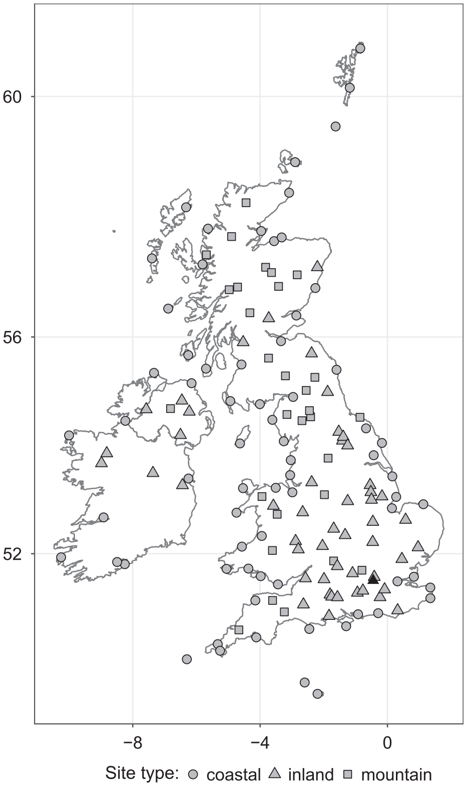


FIGURE 2 Map of the 150 observation locations in the UK and the Republic of Ireland used in this study. The sites are divided into three categories: coastal (circles), inland (triangles) and mountain (squares) sites. The black triangle marks Heathrow Airport

2.2 | Ensemble model output statistics

For all their benefits, weather forecast ensembles are usually too confident and produce underdispersed forecasts (Hamill, 2001). This means that the ensemble spread does not cover all sources of uncertainty in a given weather situation and is therefore on average too narrow. Like all weather prediction models, ensembles are also subject to a deterministic bias, depending on the model's skill in varying weather situations. To correct for the bias and the underdispersion, we first apply statistical post-processing to the raw ensemble forecasts before using the new RAFT error correction method. EMOS (Gneiting *et al.,* 2005), sometimes called non-homogeneous Gaussian regression, has successfully been applied to multiple forecast models (e.g., Kann *et al.,* 2009; Scheuerer and Büermann, 2014; Feldmann *et al.,* 2015) and is a suitable method to calibrate MOGREPS-UK forecasts.

We denote a future temperature observation for a specific location and time by Y and the corresponding ensemble forecast members by $X_1, \dots, X_{12}$. The EMOS predictive distribution of Y conditional on $X_1, \dots, X_{12}$ is then defined as a Gaussian distribution:

$$Y|X_1, \dots, X_{12} \sim \mathcal{N}\left(\mu, \sigma^2\right). \tag{1}$$

The moments of this distribution are modelled using the ensemble forecast's statistics; the predictive mean

$$\mu = a + b^2 \cdot \overline{X} \tag{2}$$

is a linear function of the ensemble mean $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ and the predictive variance

$$\sigma^2 = c^2 + d^2 \cdot S^2 \tag{3}$$

an affine function of the ensemble variance $S^2 = \frac{1}{m}\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2$. Here, $m = 12$ is the number of ensemble members and the coefficients a, b, c and d are real numbers. For estimating a, b, c and d, we use minimum score estimation (Dawid *et al.,* 2016) and optimize the continuous ranked probability score (CRPS; Matheson and Winkler, 1976; Gneiting and Raftery, 2007) based on training data as suggested by Gneiting *et al.,* (2005). Gebetsberger *et al.,* (2018) gives a comprehensive comparison of minimum CRPS and maximum likelihood estimation. The parameters in Equation (3) are squared to ensure that the predictive variance is non-negative. In Equation (2), b is constrained in the same way, making it easier to interpret.

All runs of the NWP model and all forecast lead times are calibrated separately using a rolling training period of 40 days. This means that for each run and each lead time, we collect all forecast–observation pairs from the last 40 days, where the forecasts were initialized at the same time of day and are valid for the same lead time. These data comprise the basis for the estimation of the EMOS coefficients. The current ensemble forecasts are then plugged into Equations (2) and (3) to obtain the full EMOS predictive distribution $\mathcal{N}\left(\mu, \sigma^2\right)$. We follow the local EMOS approach, in that all stations are treated on an individual basis. This accounts for local effects and turns out to produce much better results than a regional approach, where data from different sites are pooled together. In order to have a full set of training data for the first model runs in 2014, some dates from the end of 2013 are used.

2.3 | Verification methods and EMOS forecast skill

To evaluate the effectiveness of the EMOS method, we compare the predictive skill of the post-processed forecasts

TABLE 1 Continuous ranked probability score (CRPS) and root-mean-square error (RMSE) averaged over all sites and forecast runs, for different lead time ranges

	CRPS			RMSE		
Lead times	**1–12 hr**	**13–24 hr**	**25–36 hr**	**1–12 hr**	**13–24 hr**	**25–36 hr**
Raw ensemble	0.718	0.741	0.792	1.205	1.254	1.343
EMOS	0.555	0.596	0.636	1.054	1.131	1.204

The margin of error based on a 95% bootstrap interval is less than 0.002.

to the raw MOGREPS-UK ensemble. The tools used here, as well as for evaluating the RAFT forecasts in Section 4, are the root-mean-square error (RMSE), the CRPS and the rank and probability integral transform (PIT) histograms. Both the RMSE and the CRPS are proper scoring rules (Gneiting and Raftery, 2007); they measure the skill of a forecast by assigning a numerical penalty depending on how well the forecasts match the observations. It is essential that they are proper, as this guarantees that the best forecast model will receive the best score and prohibits hedging. While the RMSE assesses the deterministic forecast accuracy of the mean of a predictive probability distribution F, the CRPS evaluates the probabilistic skill of the whole distribution – which can also be represented by a discrete ensemble. The RMSE is defined as the square root of the average squared distance between the mean forecasts and the observations y:

$$\mathrm{RMSE}\,(F,y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{mean}\,(F) - y)^2}, \tag{4}$$

where n is the number of data points or forecast cases.

In its general form, the CRPS can be expressed as the squared area between a forecast cumulative distribution function (CDF) F and the empirical CDF of the observation y or, equivalently, in terms of two expected values (Thorarinsdottir and Schuhen, 2018):

$$\begin{aligned}\mathrm{CRPS}\,(F,y) &= \int_{-\infty}^{+\infty}\left[F\,(x) - \mathbb{1}\,\{y \le x\}\right]^2\,\mathrm{d}x\\ &= \mathbb{E}|X-y| - \frac{1}{2}\mathbb{E}|X-X'|,\end{aligned} \tag{5}$$

where $\mathbb{E}$ denotes the expected value with respect to F and X, X' are independent random values with distribution F. Here, we use the closed form for a Gaussian distribution (Equation (6)) to evaluate the EMOS forecasts and an approximation for the MOGREPS-UK forecasts, where the distribution is given by an ensemble (Equation (7)):

$$\begin{aligned}\mathrm{CRPS}_{\mathrm{EMOS}}\left(\mathcal{N}\left(\mu,\sigma^2\right),y\right) &= \sigma\left\{\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right)-1\right]\right.\\ &\left.\quad + 2\phi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right\},\end{aligned} \tag{6}$$

$$\begin{aligned}\mathrm{CRPS}_{\mathrm{ENS}}\,(X_1,\dots,X_m;y) =& \frac{1}{m}\sum_{i=1}^{m}|X_i-y|\\ &- \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left|X_i-X_j\right|.\end{aligned} \tag{7}$$

The functions $\Phi\,(\cdot)$ and $\phi\,(\cdot)$ in Equation (6) indicate the CDF and the probability density function (PDF) of a standard Gaussian distribution, respectively. As noted by Ferro *et al.,* (2008), the size of the ensemble may influence the $\mathrm{CRPS}_{\mathrm{ENS}}$ in Equation (7), in that larger ensembles are likely to obtain a better score. Gneiting *et al.,* (2005) gives a derivation of the result in Equation (6) and Grimit *et al.,* (2006) a derivation of the result in Equation (7).

Table 1 summarizes both scores for the EMOS and the raw MOGREPS-UK forecasts. We divide the forecast lead times into three categories, early (1 to 12 hr), mid-range (13 to 24 hr) and later lead times (25 to 36 hr) and average the scores over each of the categories. As can be expected, the scores deteriorate with increasing lead time, for both EMOS and raw ensemble forecasts. By applying the EMOS post-processing technique, the probabilistic forecast skill is improved by around 20% and the deterministic skill of the mean forecast by around 10%.

To assess the calibration of the probabilistic forecasts, we use the PIT histogram to check the level of calibration (Thorarinsdottir and Schuhen, 2018). For a perfectly calibrated forecast, the PIT values, computed by evaluating the forecast CDFs at the observations, should form a flat histogram. The equivalent method for discrete ensemble forecasts is the verification rank histogram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand *et al.,* 1997), which measures the distribution of the observation rank in the set of ensemble forecasts. Both histograms are interpreted in the same way.

In Figure 3, the PIT histograms for the EMOS forecasts are shown. Overall, they seem reasonably flat, however it seems that small miscalibrations remain; there are, in particular, too many observations that land in the lower tail of the predictive distribution. There is almost no difference in the degree of calibration for the different lead time categories. These results indicate that a major jump in forecast skill can be achieved by applying EMOS to the raw ensemble. In a next step, the forecast trajectories

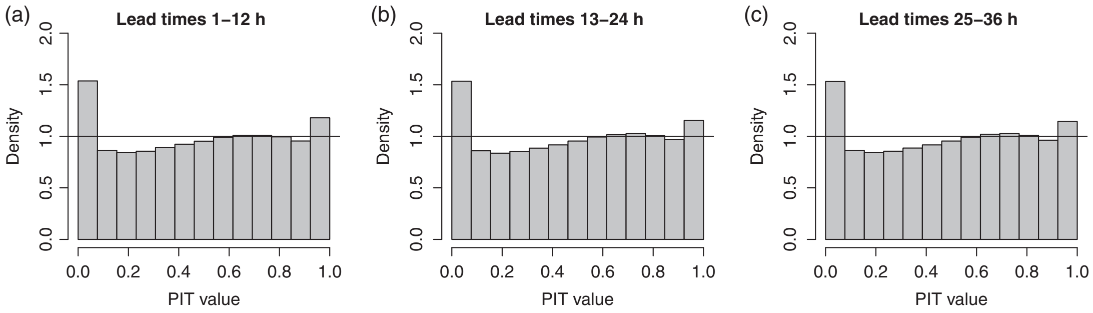


FIGURE 3 Probability integral transform (PIT) histograms of the EMOS post-processed forecasts, indicating the degree of calibration. Forecast cases are aggregated over all sites and forecast runs in the test set for (a) early, (b) mid-range and (c) later lead times

provided by the EMOS mean are successively updated using the RAFT technique. Therefore, EMOS forms a baseline against which all further error reduction is measured.

3 | RAPID ADJUSTMENT OF FORECAST TRAJECTORIES

The new RAFT technique is applied directly to the mean of the EMOS forecast distribution, in order to increase the deterministic skill of the EMOS forecasts even further when new information becomes available. This in turn also leads to a reduction in the CRPS (Equation (5)). More specifically, the goal of the new RAFT method is to adjust and improve forecast trajectories over time by using the part of the trajectory that has already verified, in conjunction with the matching observations. First we need to establish the relationship between forecast errors at different lead times. The forecast error $e_{t,l}$ is here defined as the distance of the EMOS mean forecast $\mu_{t,l}$ to the observation y_{t+l}, where the forecast is initialized at time t and valid at lead time l:

$$e_{t,l} = y_{t+l} - \mu_{t,l}. \tag{8}$$

Figure 4a shows the Pearson correlation coefficient matrix of the forecast errors at Heathrow Airport (marked with a black triangle in Figure 2) for the 0300 UTC model run. To create the plot, the error correlations for all possible pairs of lead times were computed over the training set, as well as the corresponding p-values. Only statistically significant correlations at the 90% level are shown. The correlation between lead times 1 and 36 is slightly negative and significant, but is left out for clarity and ultimately has no relevance for this study.

In all instances, there is a positive correlation between the errors at a certain lead time and its immediate neighbours. This means that the errors at two lead times, if close enough, are so strongly connected that we can make inference about the forecast skill at a future lead time by observing the error at the earlier lead time. Formally, there is a period preceding each forecast $\mu_{t,l}$, during which the recently measured forecast error e_{t,l^*}, with $l^* < l$, provides useful information for a forecast adjustment at time $t + l$ and thus can reduce the subsequent error $e_{t,l}$.

The size of these temporal neighbourhoods varies greatly with the time of day. At lead times 8 to 11, corresponding to midday, the relationship between the forecasts is weakest with only 4 to 5 hr of significant correlation, while the largest predictability of 15 to even 27 hr can be found at lead times 28 to 31, in the early morning. In the MOGREPS-UK setting, this makes the RAFT method work on a rather short time-scale, adjusting forecasts sometimes at only a couple of hours in advance. However, RAFT adapts to the scale and context of the application; for example, for daily weather forecasts, the potential time range of adjustment increases to a few days.

Based on the correlation structure in Figure 4a, we can now define the RAFT model, establishing the relationship between forecast errors at two different lead times by linear regression. The estimated future error $\hat{e}_l$ at lead time $l = 1, \dots, 36$ is written as a linear function of the observed error at earlier lead times l^*:

$$\hat{e}_l = \hat{\alpha} + \hat{\beta} \cdot e_{l^*} + \varepsilon. \tag{9}$$

The error term ε is normally distributed with mean zero and both coefficient estimates $\hat{\alpha}$ and $\hat{\beta}$ are determined by the least squares approach based on the training dataset. All lead time combinations, sites and the four NWP model initialization times are treated separately. We omit the index t for the model run from Equation (9) for simplicity. Regression equations using multiple lead times as predictors were also investigated, but did not yield any improvement, as the newest observation always contains the most useful information.

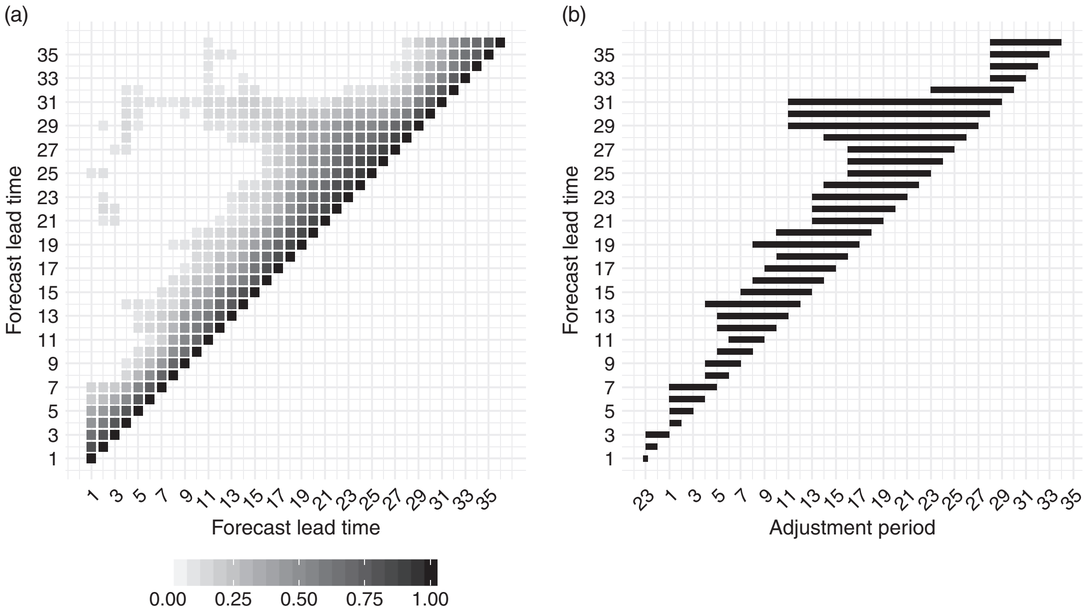


FIGURE 4 (a) Empirical correlation coefficient of the forecast error for every lead time combination of the 0300 UTC model run at Heathrow Airport during the training period from January to December 2014. Correlations are only plotted if they are significant at the 90% level. (b) RAFT adjustment period for each forecast lead time for the 0300 UTC model run at Heathrow Airport. The periods refer to the time points at which the observations used to adjust the future forecast are recorded

Not all of the possible lead time combinations produce valid and useful results. As seen in Figure 4a, the correlation between lead times, and therefore predictability, is irregular and depends on various factors. Consequently, we define for each lead time l an adjustment period of length p, consisting of the preceding lead times for which there is a strong enough correlation to affect the forecast skill. Starting at $l-p+1$, the forecast for lead time l is repeatedly adjusted in hourly steps, each time using the most recent available forecast error information. Here, we allow for a processing time of one hour after an observation has been recorded, which means that the final adjustment for a forecast valid at lead time l is made at $l-1$, using the error at $l-2$.

To establish the length of the adjustment period for each location and lead time numerically, we need an algorithm that ensures that any adjustments are not based on random effects, but genuine additional error information about future lead times. Therefore we look at the coefficient $\hat{\beta}$ in the RAFT model (Equation (9)) and determine for which lead time combinations the estimate is significantly different from zero. This corresponds to a large enough error correlation between the two lead times at hand to justify a RAFT adjustment. As to the level of significance for $\hat{\beta}$, we want to be a little lenient if the temporal difference between lead times is small, starting with a level of 90%, and become stricter with increasing distance to the predicted lead time, ending with 99%. Experiments have shown that legitimate connections at small lead time differences can be missed if the required level is set too high and spurious correlations at far apart lead times can lead to excessively long adjustment periods without real improvement if the level is set too low.

With our forecast trajectories spanning 36 hr, we need to account for the fact that multiple lead times correspond to the same time of day. As we treat each of the four forecast runs individually, there is for every lead time combination separated by more than 24 hr a corresponding combination from the run initialized one day earlier with a time difference of less than 24 hr, which will on average provide a more skilful forecast. Therefore, the maximum length of the adjustment period is 22 hr, with one hour allowed for processing the observations.

The algorithm for obtaining the optimal length of the adjustment period is then defined as follows:

1. Run the linear regression in Equation (9) using all lead times $l^* \in [l-23, l-2]$ as predictors. For negative lead times, add 24 hours, so that lead time 23 is followed by lead time 0, 1, etc.

2(a) Working backwards, find the first instance of l^* in $[l-11; l-2]$ where the regression coefficient $\hat{\beta}$ is not significantly different from zero at the 90% level. If a result can be found, we denote it by l_p.
(b) If such an l_p cannot be found, find the first instance of l^* in $[l-19; l-12]$ where $\hat{\beta}$ is not significantly different from zero at the 95% level. If a result can be found, we denote it by l_p.
(c) If such an l_p cannot be found, find the first instance of l^* in $[l-23; l-20]$ where $\hat{\beta}$ is not significantly different from zero at the 99% level.

3. Set $p = l - l_p$. If no value for l_P is found after Step 2, then p is the average of the adjustment period lengths of the neighbouring lead times $l-1$ and $l+1$. In case this does not produce a valid number, p is set to 22, the maximum possible length for the adjustment period.

This somewhat arbitrary algorithm was designed so that it works well for a multitude of sites in our dataset with very different correlation patterns. It can be replaced by any other method for identifying a suitable adjustment period. Figure 4b shows the adjustment periods for the 0300 UTC run at Heathrow Airport produced by the algorithm above. It is clear that there is a strong connection between the correlation pattern in Figure 4a and the adjustment period length, in that large p correspond to longer periods of predictability. Note that for a stable estimation, the algorithm is applied only once to the entire training set (data from January to December 2014) with the obtained parameter estimates used for all data in the test set, as opposed to the rolling training period approach used for the EMOS post-processing.

The adjustment period refers to the time points when the observations used in the adjustment are recorded, and not the time points when the adjustments are carried out. As we allow an extra hour for the processing of the observations, the actual correction is made one hour after the observation time, starting at $l - p + 1$. For example, we see from Figure 4b that the ideal length of the adjustment period for lead time $l = 25$ here is $p = 9$. This means that the first correction to a forecast valid at $t + 25$ is made at $t + l - p + 1 = t + 17$ using the observation collected 1 hr earlier, at $t + 16$. From there on, an adjustment takes place every hour, each time using the newest error information available at that moment, until the time $t + 24$, where we adjust the forecast for a final time based on the error measured at $t + 23$. Clearly this last observation gives us the most accurate information about the expected forecast error, as it is closest in time to the forecast. This means that we get the most gain in forecast skill if RAFT is applied in the very short term.

Obviously, there is a gap during the first 2 hr of the forecast trajectory, where no forecast data from the current run are available to adjust the forecasts at $t + 1$ and $t + 2$. In this case, we instead use forecasts from the run that was initialized 24 hr earlier which are valid at the same time as the missing forecasts. Of course this does not lead to the same kind of improvement in forecast skill, as the current forecast run might exhibit a very different error characteristic from the one from 24 hr ago.

To obtain the size and direction of the forecast adjustment for a certain forecast run t and lead time l, we first calculate the observed error $e_{t,l-k}$ at lead time $l - k$ according to Equation (8), where $k \leq p$ and the time $l - k$ thus lies within the adjustment period. Then we plug the observed error into the regression equation for the predicted error $\hat{e}_{t,l}$ at the future time point $t + l$:

$$\hat{e}_{t,l} = \hat{\alpha} + \hat{\beta} \cdot e_{t,l-k}. \tag{10}$$

The regression coefficient estimates $\hat{\alpha}$ and $\hat{\beta}$ are unique for each lead time combination, forecast initialization time and location, and were calculated in the first step of the algorithm to find the optimal adjustment period. Once we have established the predicted error in this way, we add it to the EMOS mean forecast $\mu_{t,l}$ and obtain the adjusted RAFT forecast $\hat{\mu}_{t,l}$:

$$\hat{\mu}_{t,l} = \mu_{t,l} + \hat{e}_{t,l}. \tag{11}$$

The resulting adjusted mean forecast is generated from data that have passed through multiple levels of post-processing. First, while applying EMOS, the performance of the raw ensemble over the past 40 days is analyzed and the results are used to improve the deterministic and probabilistic forecast skill. This post-processing method uses forecasts and observations from a rolling training period and is carried out right after the NWP model run has finished and before the forecast is issued. When the first forecast from the trajectory verifies 2 hr later, we make the first RAFT adjustment and continue in the same manner in hourly intervals (Figure 5). The level of RAFT error correction only relies on the performance of the EMOS forecast mean during the current forecast run, using very short-term information not available when the NWP model was initialized and when EMOS was applied. The combined EMOS/RAFT predictive distribution consisting of the RAFT forecast as mean and the EMOS variance can produce a more accurate forecast than both the raw ensemble and the unadjusted EMOS forecast, while remaining calibrated.

4 | RESULTS

In the previous section, we described how the RAFT method can be combined with post-processing methods

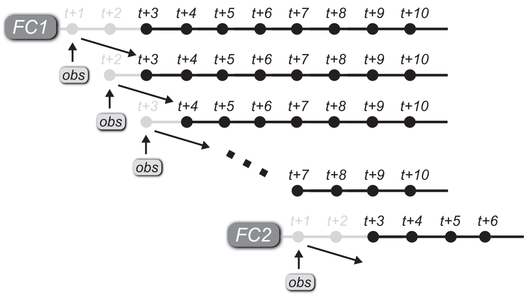


FIGURE 5 Diagram of a forecast cycle for an hourly forecast issued every 6 hr with rapid adjustment of the forecast trajectory (RAFT) applied as new observations become available. Forecasts in grey are only used as predictors by means of their observed error and are not adjusted themselves

like EMOS to provide an additional short-term error correction. We now show comprehensive results, first for Heathrow Airport and then for all sites in the dataset.

4.1 | Results for Heathrow Airport

As the busiest airport in the UK, accurate weather forecast for Heathrow are of major importance, especially for the very short term (e.g., Ghirardelli and Glahn, 2010). Therefore, we investigate the impact of RAFT on forecast quality at this site separately. From Figure 4, we know how the relationship between forecast errors at different lead times can be used to define the RAFT regression model and corresponding adjustment periods. This analysis is done only once and the parameters are then valid until there are significant changes in the forecast models or the local error characteristics.

In the following example, we illustrate how RAFT works in a real-time setting. Figure 6 is a snapshot, taken at 2300 UTC on 14 January 2016 at Heathrow Airport. The light grey dashed line depicts a forecast trajectory, issued at 0300 UTC the same day and post-processed using EMOS as described in Section 2.2. Over time, temperature values (represented by the black solid line) are observed for the 36 lead times of the trajectory. However, at the time of the snapshot, they are only available up to 1 hr before. The dot-dashed dark grey line is the RAFT forecast and consists of two parts. The trajectory left of the black vertical line is a combination of the most recent RAFT forecasts at each lead time, i.e., the forecast issued 1 hr earlier, using the error information from 2 hr before the valid time. These are the optimal RAFT forecasts, as they contain the most information and are very short-term.

The right side of the black vertical line is the current RAFT trajectory, showing the best possible forecast we can make with the information we have at this point in time. Depending on the length of the adjustment period, the forecasts from here to the end of the original forecast trajectory are adjusted using the most recent error information. For example, the forecast at $t + 28$ is being adjusted, while the forecast at $t + 33$ is not. For the first 12 hr, the uncorrected trajectory has a good agreement with the observations and only small corrections are made. Between lead times 15 and 30, corresponding to evening and night-time, the EMOS forecast underpredicts the temperature. As soon as larger errors are observed, the RAFT adjustment to the original forecast also becomes larger and after a short time manages to counter the underprediction. This example illustrates how RAFT is able to quickly correct forecast errors a few hours ahead, whereas the unadjusted forecast would continue to underpredict the temperature for further 15 hr.

To evaluate the performance of RAFT over the entire test period, we look at the root-mean-square error of the RAFT-adjusted forecasts and compare to the unadjusted EMOS mean forecasts. Figure 7 shows the RMSE at Heathrow, averaged over all cases in the test period where the NWP model was initialized at 0300 UTC. In both plots, the solid line is identical and represents the performance of the EMOS-post-processed forecast trajectories, and the dashed line is the RMSE of the RAFT forecasts. The difference between the plots lies in the fact that they are snapshots taken at different points in the forecast cycle.

Figure 7a depicts the level of forecast skill if we stopped applying RAFT after lead time $t + 15$. This would mean that all forecasts to the left of the vertical line have been adjusted according to the forecast error measured 2 hr earlier. As the most recent observed error is registered at $t + 14$, all forecasts to the right of the vertical line are adjusted using this error information (depending on the length of the respective adjustment periods). This means that on the left side, the difference between the two curves is the maximum improvement obtainable by applying RAFT.

For the first few hours, there is only very little improvement, as we do not yet have any information about the current run's forecast error, and we have to rely on the information from the run started 24 hr earlier. However, as soon as the new error information is available, RAFT shows a considerable reduction in forecast error, even up to 20%. On the right side, the largest benefit can be seen in the next few hours, as the correlation is strongest between close lead times. After about 5 hr, RAFT falls back to the skill level of the EMOS forecasts. Interestingly, for the period between $t + 28$ and $t + 32$, there appears to be a significant correlation to the error at $t + 14$. Thus we see a small error reduction 14 to 18 hr ahead.

In Figure 7b, a different snapshot is shown. Now we apply RAFT to the full forecast cycle, that is, we let it run

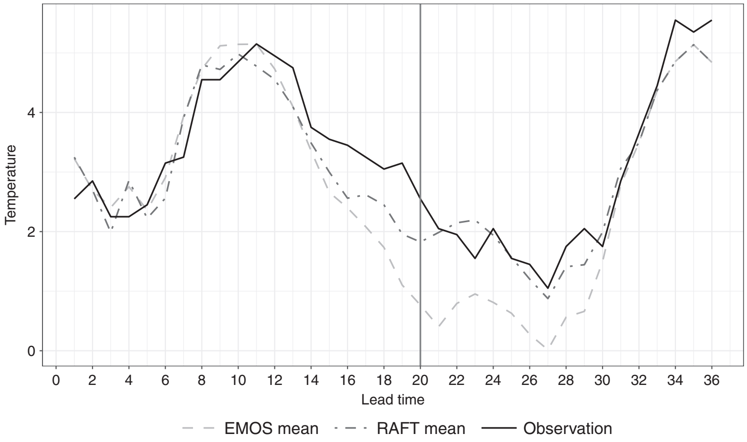


FIGURE 6 Example forecast at Heathrow Airport. Snapshot of the RAFT and EMOS trajectories taken at 2300 UTC on 14 January 2016 (denoted by the vertical line), where the model was initialized 20 hr earlier. See the text for further details

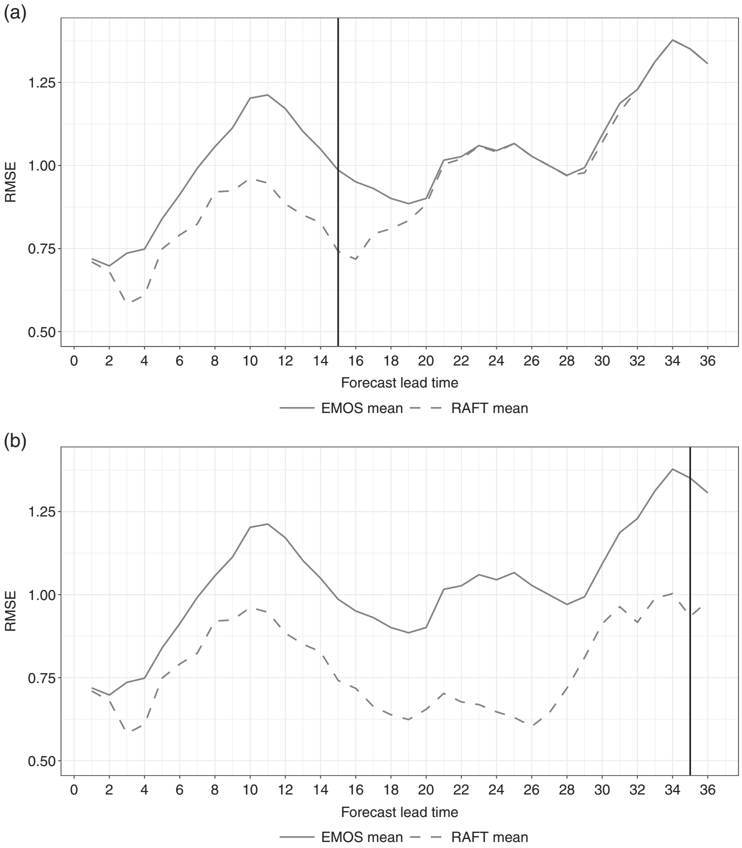


FIGURE 7 (a) RMSE of the EMOS and RAFT mean forecasts over lead time. The scores are averaged over all dates in the test period at Heathrow Airport for model runs initialized at 0300 UTC. RAFT error corrections are only carried out until lead time $t+15$. (b) is as (a), but RAFT is carried out for all lead times until the end of the trajectory

until the last adjustment made at $t+35$. This plot considers only the most short-term correction for each lead time and therefore the best possible forecast. Here we see a large improvement over EMOS throughout and especially for later lead times. An interesting feature emerges if we compare the forecast skill at lead times $t+2$ and $t+26$. These lead times correspond to the same time of day, 0500 UTC, and we would expect the forecasts at $t+26$ to

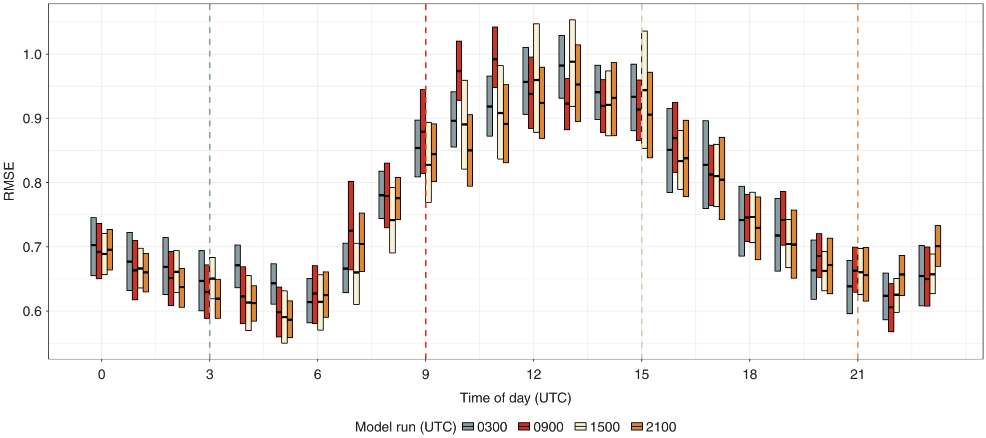


FIGURE 8 Average root-mean-square error for RAFT forecasts from all four daily NWP model runs as a function of the time of day. The scores were computed over the test period at Heathrow Airport and are shown with 90% bootstrap confidence intervals. The dashed vertical lines represent the initialization times of the NWP model

perform worse due to more time having passed since the model initialization. With RAFT, however, this forecast was adjusted with a very recently measured observation error, whereas the $t + 2$ forecast could only be adjusted using the data from the model run initialized 24 hr prior. As a result, the $t + 26$ error is lower than the one at $t + 2$ and, consequently, a forecast for $t + 26$ of an older model run will on average have more forecast skill than the $t + 2$ forecast from the next (and newer) model run.

This means that there is a transition period at the beginning of every NWP model run, where an old run provides better forecasts until the point is reached where the forecasts from the new model run can be used for the RAFT adjustment. Figure 8 illustrates the relationship between all four initialization times, depicting the average RAFT RMSE as a function of the time of day in UTC. The times when a new model run is started are marked by dashed vertical lines. Again, the RMSE is computed using the most recent adjusted and optimal forecast. Here, the mean score is shown, as well as 90% confidence intervals based on 1,000 bootstrap samples.

At first glance, there is a strong diurnal variation in all four runs, with the lowest predictability around midday and the highest during the early morning. We are interested in the ranking of the four runs in terms of forecast skill. Ordinarily, we would expect the newest run to be the best, but as seen in Figure 7b, there is a short period during which an older run produces better forecasts. For the first few hours of the day, the ranking is as expected, in that the 2100 UTC run has the lowest RMSE and the 0300 UTC run the highest. When the first forecast from the new 0300 UTC run comes in at 0400 UTC, the skill decreases considerably, instead of improving. This is due to the fact that there are no recent forecast data available for the RAFT adjustment and we have to rely on the error information from 24 hr before. For 2 hr after the initialization of the 0300 UTC run, the 2100 UTC run remains the best forecast; the score difference between the two runs is actually significant at the 90% level. After 0600 UTC, the model runs rank in the expected order.

A similar pattern can be noticed every time a new model run is produced, with the exception being the 1500 UTC run. This run actually ranks best, or at least close to the others, from the first forecast, coinciding with the increase in predictability in the afternoon. We can conclude that the four daily model runs have comparable forecast quality after applying RAFT, apart from a transition period of about 2 hr. During this period, forecasts from an older run should be preferred to the newest.

4.2 | Results for all sites

After presenting the results for Heathrow Airport, we now discuss how RAFT performs for all observation sites available. The dataset covers the British Isles (Figure 2) and displays a wide variety of local characteristics, such as sites in the Scottish mountains at elevations above 1000 m or coastal towns.

Figures 9a,b compare the average RMSE of the EMOS and RAFT forecasts for the 2100 UTC model run, similar to Figure 7. Again, they represent snapshots at different times in the RAFT adjustment process. In Figure 9a, we see the maximum achievable RAFT improvement over the

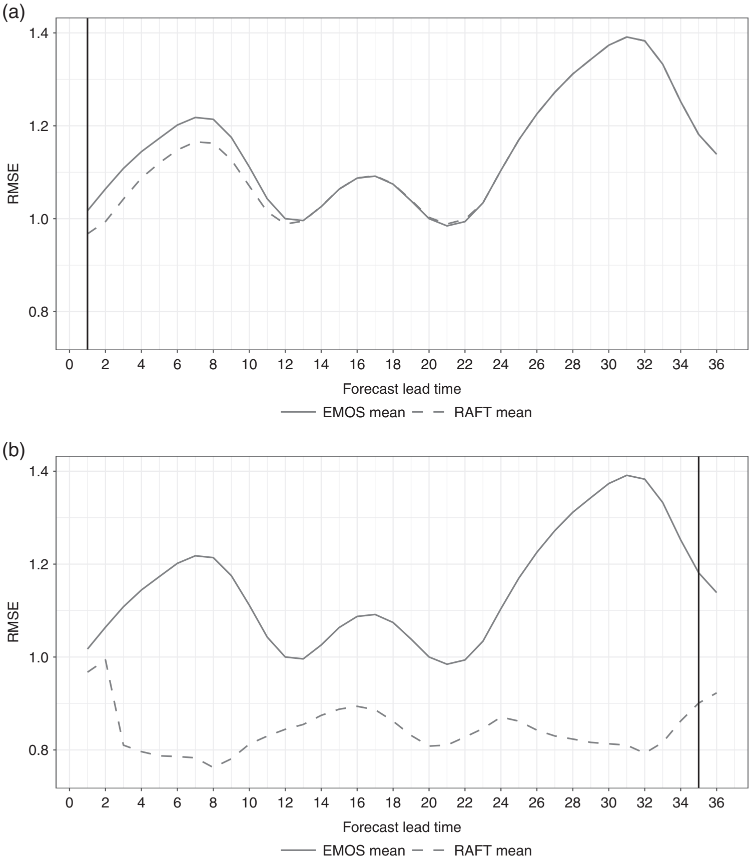


FIGURE 9 (a) RMSE of the EMOS and RAFT mean forecasts over lead time. The scores are averaged over all dates and locations in the test period for model runs initialized at 2100 UTC. RAFT error corrections are carried out only once at lead time $t + 1$. (b) is as (a), but RAFT is carried out for all lead times until the end of the trajectory

EMOS mean if we only applied the adjustment once at the moment the first forecast becomes valid at $t + 1$. At that time, no observations are available yet for the new run, so we have to rely solely on error information from the run initialized 24 hr earlier. Those RAFT forecasts for which the adjustment period extends beyond the beginning of the run have been adjusted using the observation made at $t +$ 0, combined with the old run's $t + 24$ forecast.

While the benefit from applying RAFT in this way is considerably smaller than the improvement we see as soon as the new forecast data are used, there is still a reduction in the RMSE for the next 12 hr. We notice an interesting detail between $t + 20$ and $t + 23$ (corresponding to 1700 UTC and 2000 UTC, respectively). In this period of high predictability, the RAFT scores are actually slightly worse than the EMOS scores, but revert to being equal with the next RAFT adjustment at $t + 2$ (not shown). This pattern can be observed at a handful of sites, where the error correlation between the lead times is particularly strong and the corresponding adjustment periods quite long. The RAFT algorithm described in Section 3 is applied in the same form to all locations and lead times. This does not take into account any potential stark differences in correlation patterns between the sites which in turn might require slightly different stopping rules or significance levels for an optimal performance. It might therefore be advisable to look into adjusting the algorithm if interest is in optimizing the performance for specific locations.

In Figure 9b, we again see the outcome if RAFT is applied every hour up until the last installment at $t + 35$. This represents the maximum and most short-term gain in forecast skill achievable at every lead time and is not a continuous trajectory. We will use these forecasts for the entire subsequent analysis. At the beginning of the forecast cycle, there is a sharp drop in the RMSE, immediately after we are able to use data from the current run. Afterwards, the RAFT skill remains relatively constant, with small variations due to the diurnal cycle, whereas the EMOS skill fluctuates considerably. Especially during the last 12 hr of the forecast cycle, the improvement of RAFT over EMOS is

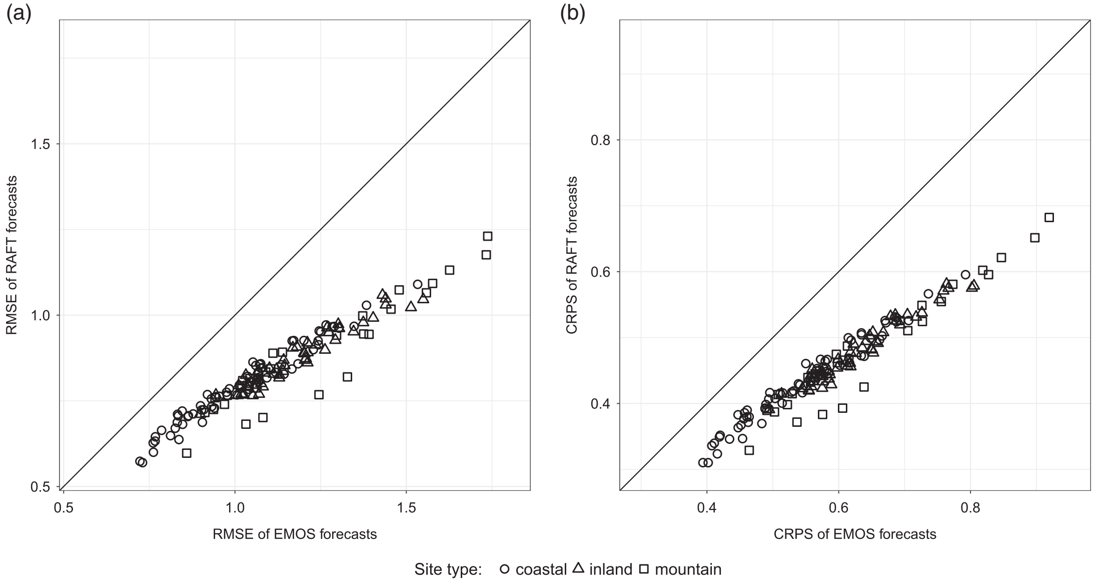


FIGURE 10 (a) Average RMSE of the RAFT forecasts as function of the EMOS RMSE for all sites, lead times and model runs in the dataset. (b) Average CRPS of the RAFT forecasts as function of the EMOS CRPS for all sites, lead times and model runs in the dataset. The RAFT CRPS is computed using the EMOS predictive variance

quite substantial, as the short-term RAFT forecast corrections manage to cancel out the skill deterioration usually occurring with increasing lead time.

All observation sites in the study can be separated into three categories based on their location: coastal, inland and mountain sites. In Figure 10, the RMSE and CRPS scores for all locations are aggregated over all four model runs and the RAFT scores are plotted against the EMOS scores. The CRPS for RAFT is calculated by plugging the RAFT mean into the EMOS predictive distribution. For both the RMSE and the CRPS, we see an improvement for all sites after applying RAFT, in particular at locations where the error was high in the first place. In fact, the improvements seem to follow the same linear trend, apart from a group of five mountain sites (located in Scotland and Cumbria), which receive a somewhat larger benefit from RAFT than the other sites. This hints at some location-specific issues not resolved by EMOS or the original ensemble.

In Figure 3, we showed that EMOS produces nearly calibrated forecasts and naturally we want to preserve this level of calibration with RAFT. Therefore we compare the rank and PIT histograms of the raw ensemble, EMOS and the distribution consisting of the RAFT mean and the EMOS predictive variance. Figure 11 shows these histograms divided by site type. For all three forecasting methods, there is only very little difference in calibration between coastal, inland and mountain sites. The raw ensemble is, as expected, uncalibrated and very underdispersive, recognizable by the characteristic U-shape. EMOS is fairly calibrated, although there is still some hint of a bias and underdispersion. In contrast, RAFT is slightly overdispersive, meaning that the variance of the distribution is on average too large. This is not surprising, given that the mean of the distribution now has much better deterministic skill, but the corresponding EMOS variance has not changed. An additional adjustment of the EMOS variance to counteract the induced overdispersion is a potential subject for further study.

Another indicator of calibration is the actual coverage of the prediction interval compared to the nominal value. The ensemble members create a prediction interval of $11/13 \approx 84.62\%$, which would correspond to perfect calibration. However, the raw MOGREPS-UK ensemble only reaches a coverage of 52.24%, whereas the EMOS coverage is 79.29% and the RAFT prediction intervals cover 87.31%. Although one is under- and the other overdispersive, both EMOS and RAFT are nearly calibrated, with the coverage for RAFT being slightly closer to the correct value.

Finally, we look at how RAFT performs during different seasons of the year. The test set contains two full spring seasons, and one full winter, summer and autumn. Figure 12 depicts the RMSE skill score, the relative improvement of the RAFT over the EMOS mean,

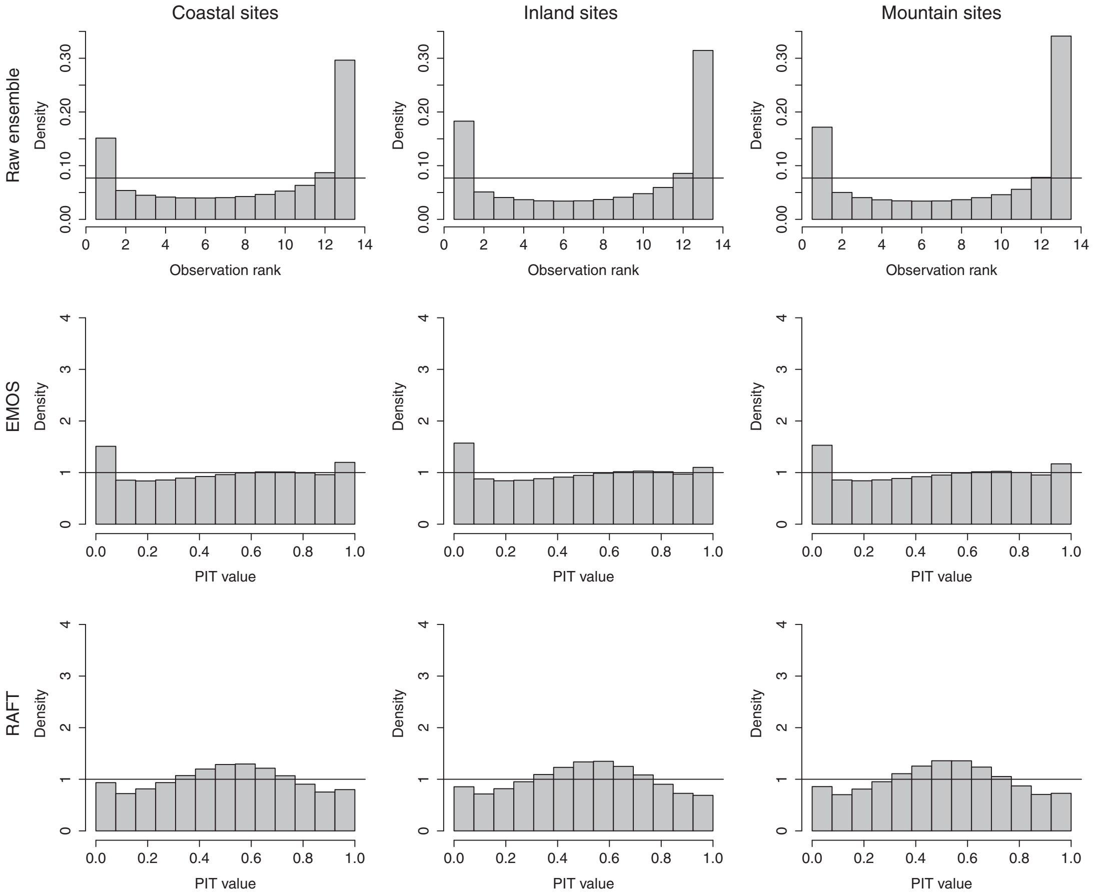


FIGURE 11 Verification rank histograms for the raw ensemble (top row) and PIT histograms for the EMOS (middle row) and RAFT (bottom row) forecasts. The RAFT predictive distribution is generated by using the EMOS predictive variance. The histograms are divided by site type and data are aggregated over all dates, lead times and model runs in the test dataset

for the four seasons. A score of 1 would mean a perfect forecast and a score of 0 no improvement over the reference forecast. Again, all four runs and all sites have been aggregated.

The largest gain in forecast skill occurs during the night and is very similar for all seasons. The same pattern holds for the time between 1200 and 1600 UTC, where the skill score values are very close. In the morning, however, the scores for summer and winter behave very differently; they both decrease, but the summer skill score much faster and further than the winter score. This is due to the fact that in summer, the diurnal cycle plays a much more prominent role (not shown) and the predictability during night is much higher than during the day. In winter, the RMSE is more stable and there is only very little difference in predictability. The deterioration in the skill score during the early morning in summer coincides with a period of large change in predictability. It seems that during this time predictability changes so fast that even the very short-term RAFT adjustment can only improve the forecast skill by a small amount. Therefore, it might be advantageous to look into obtaining separate RAFT coefficients for the different seasons. This is not possible in the context of the current study, however, as a much larger training dataset would be required.

5 | CONCLUSIONS AND DISCUSSION

This paper presents a new post-processing approach for NWP forecasts, rapid adjustment of forecast trajectories

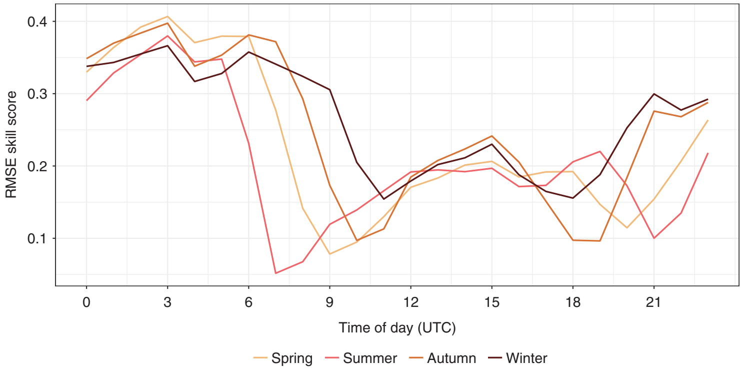


FIGURE 12 RMSE skill score of RAFT with EMOS as reference forecast against time of day for different seasons. RMSE scores are averaged over all sites, model runs and dates in the test dataset

(RAFT), which is applied on top of the traditional post-processing approach EMOS once new information pertaining to the current forecast trajectory becomes available. By utilizing the forecast error correlation structure in the post-processed NWP forecast trajectories, the EMOS mean forecasts of the not-yet-realized part of the trajectory are adjusted in every time step of the forecast based on the forecast errors that have already been realized. This computationally efficient approach to make use of the newest available information provides an appealing alternative to computationally costly rapid ensemble cycles (Lu *et al.,* 2007; Benjamin *et al.,* 2016), and the older forecast gains skill in the time between initialization and release of the next NWP forecast cycle.

While the precise set-up described here may have some operational restrictions due to computing and observation processing time if applied at a large number of locations, our results provide a convincing proof-of-concept. For example, as shown in Figure 9b, the forecast skill may be improved by over 40% on average in terms of RMSE when a 32-hour-old forecast is supplemented with the most recent available information an hour before it is realized. In an operational setting, the amount of benefit from the RAFT approach will depend heavily on the operational set-up of the forecast system. The MOGREPS-UK data used here were run on a 6-hourly basis, which is quite typical for a NWP system. For this type of set-up, our results at Heathrow Airport suggest a potential new strategy for updating the forecast cycle in that a delay in introducing the new NWP forecast may be preferred if RAFT is employed. Since spring 2019, MOGREPS-UK has changed to run on an hourly-updating cycle, with three members run every hour and an 18-member ensemble formed by time-lagging of six cycles. In such cases, it might be beneficial to apply RAFT to the older members of the time-lagged ensemble; Schuhen (2019) gives an application of RAFT to individual ensemble members.

RAFT is easily implemented at individual locations and could be especially useful for forecast users in applications such as aviation and renewable energy production where decision-making relies on location-specific skilful weather forecasts. Here, the forecast user commonly has access to their own observations in close to real time while the NWP forecast may be delivered with a small time lag, or a decision needs to be made in the middle of a forecast cycle, making the setting ideal for a RAFT application. In such cases, observation frequency may also be higher than the time resolution of the NWP forecast, a situation to which RAFT can easily be adapted.

In the EMOS post-processing procedure, each lead time is corrected independently based on forecast errors pertaining to that same lead time in older forecasts. As noted by (e.g.) Schefzik *et al.,* (2013), this may lead to physical inconsistencies between lead times so that the EMOS mean trajectory over all lead times may not be a physically consistent forecast trajectory. One potential inconsistency is unrealistically large jumps in the temperature between lead times. Using the convergence index proposed by Ehret (2010), we compared the temporal stability, or the jumpiness, of the EMOS mean trajectory and the last RAFT trajectory and found the RAFT trajectory to be less jumpy for almost all sites than the original EMOS mean trajectory. This indicates that RAFT might correct some of the physical inconsistencies across lead times introduced in the univariate EMOS post-processing. An approach that combines RAFT with the ensemble copula coupling (ECC) approach of Schefzik *et al.,* (2013) to generate physically consistent trajectories for wind forecasts is proposed in Schuhen (2019).

In our analysis, we update only the mean of the EMOS forecasts while the variance remains unchanged. The original EMOS forecasts are slightly underdispersive and biased; a similar effect has been reported in previous applications of EMOS to individual locations (e.g., Thorarinsdottir and Gneiting 2010). The RAFT procedure reduces the bias and improves the overall calibration, while changing the sign of the miscalibration to slightly overdispersive, cf. Figure 11. This effect is robust across all lead times as the EMOS forecast uncertainty is nearly constant across the relatively short lead times of 1–36 hr, except for minor diurnal differences related to the diurnal predictability pattern displayed in Figure 8. Our experiments to update the EMOS spread simultaneously with the mean were not successful in that they did not result in further skill improvement. One potential explanation for this is the consistency of the EMOS spread across the lead times; as the EMOS spread for 1 hr ahead forecasts is similar to that for 36 hr ahead forecasts, we do not necessarily expect to the be able to improve upon the spread for the 36 hr ahead forecasts, even if their means are updated to become 1 hr ahead predictions. However, a joint approach for mean and spread might be worth investigating further in cases where the originally post-processed forecast is nearly perfectly calibrated, or slightly overdispersive.

ACKNOWLEDGEMENTS

The authors are grateful for the support of the Research Council of Norway, N.S. through grant no. 259864, A.L. through grant no. 237718 and T.L.T. through grant no. 270733. We thank the Met Office, in particular the IMPROVER team, for giving us the possibility to work with MOGREPS-UK data.

ORCID

N. Schuhen https://orcid.org/0000-0002-5108-9047

REFERENCES

Anderson, J.L. (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9, 1518–1530. https://doi.org/10.1175/1520-0442(1996)009<1518:amfpae>2.0.co;2

Benjamin, S.G., Weygandt, S.S., Brown, J.M., Hu, M., Alexander, C.R., Smirnova, T.G., Olson, J.B., James, E.P., Dowell, D.C., Grell, G.A., Lin, H., Peckham, S.E., Smith, T.L., Moninger, W.R., Kenyon, J.S. and Manikin, G.S. (2016) A North American hourly assimilation and model forecast cycle: the rapid refresh. *Monthly Weather Review*, 144(4), 1669–1694. https://doi.org/10.1175/mwr-d-15-0242.1

Berrocal, V.J., Raftery, A.E. and Gneiting, T. (2008) Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, 2, 1170–1193. https://doi.org/10.1214/08-AOAS203

Dawid, A.P., Musio, M. and Ventura, L. (2016) Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43, 123–138. https://doi.org/10.1111/sjos.12168

Ehret, U. (2010) Convergence index: a new performance measure for the temporal stability of operational rainfall forecasts. *Meteorologische Zeitschrift*, 19(5), 441–451. https://doi.org/10.1127/0941-2948/2010/0480

Feldmann, K., Scheuerer, M. and Thorarinsdottir, T.L. (2015) Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review*, 143(3), 955–971. https://doi.org/10.1175/mwr-d-14-00210.1

Ferro, C.A., Richardson, D.S. and Weigel, A.P. (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15(1), 19–24. https://doi.org/10.1002/met.45

Gebetsberger, M., Messner, J.W., Mayr, G.J. and Zeileis, A. (2018) Estimation methods for nonhomogeneous regression models: minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12). https://doi.org/10.1175/MWR-D-17-0364.1

Ghirardelli, J.E. and Glahn, B. (2010) The meteorological development laboratory's aviation weather prediction system. *Weather and Forecasting*, 25(4), 1027–1051. https://doi.org/10.1175/2010waf2222312.1

Glahn, H.R. and Lowry, D.A. (1972) The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, 11(8), 1203–1211. https://doi.org/10.1175/1520-0450(1972)011<1203:tuomos>2.0.co;2

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69, 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gneiting, T., Raftery, A.E., Westveld, A. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5), 1098–1118. https://doi.org/10.1175/mwr2904.1

Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378. https://doi.org/10.1198/016214506000001437

Grimit, E.P., Gneiting, T., Berrocal, V.J. and Johnson, N.A. (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132, 2925–2942. https://doi.org/10.1256/qj.05.235

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. and Tennant, W. (2017) The Met Office convective-scale ensemble, MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society*, 143, 2846–2861. https://doi.org/10.1002/qj.3135

Hamill, T.M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:iorhfv>2.0.co;2

Hamill, T.M. and Colucci, S.J. (1997) Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, 125, 1312–1327. https://doi.org/10.1175/1520-0493(1997)125<1312:voersr>2.0.co;2

Hemri, S., Lisniak, D. and Klein, B. (2015) Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, 51, 7436–7451. https://doi.org/10.1002/2014wr016473

Kann, A., Wittmann, C., Wang, Y. and Ma, X. (2009) Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Monthly Weather Review*, 137, 3373–3387. https://doi.org/10.1175/2009mwr2793.1

Lu, C., Yuan, H., Schwartz, B.E. and Benjamin, S.G. (2007) Short-range numerical weather prediction using time-lagged ensembles. *Weather and Forecasting*, 22(3), 580–595. https://doi.org/10.1175/waf999.1

Matheson, J.E. and Winkler, R.L. (1976) Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Mitchell, H.L. and Houtekamer, P.L. (2000) An adaptive ensemble Kalman filter. *Monthly Weather Review*, 128(2), 416–433. https://doi.org/10.1175/1520-0493(2000)128<0416:aaekf>2.0.co;2

Möller, A., Lenkoski, A. and Thorarinsdottir, T.L. (2013) Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139, 982–991. https://doi.org/10.1002/qj.2009

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174. https://doi.org/10.1175/MWR2906.1

Schefzik, R., Thorarinsdottir, T.L. and Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4), 616–640. https://doi.org/10.1214/13-STS443

Scheuerer, M. and Büermann, L. (2014) Spatially adaptive post-processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(3), 405–422. https://doi.org/10.1111/rssc.12040

Schuhen, N. (2019) Order of operation for multi-stage post-processing of ensemble wind forecast trajectories. *Nonlinear Processes in Geophysics Discussion*. https://doi.org/10.5194/npg-2019-55

Schuhen, N., Thorarinsdottir, T.L. and Gneiting, T. (2012) Ensemble model output statistics for wind vectors. *Monthly Weather Review*, 140, 3204–3219. https://doi.org/10.1175/mwr-d-12-00028.1

Sloughter, J.M., Gneiting, T. and Raftery, A.E. (2013) Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review*, 141(6), 2107–2119. https://doi.org/10.1175/MWR-D-12-00002.1

Talagrand, O., Vautard, R. and Strauss, B. (1997). Evaluation of probabilistic prediction systems, In Proceedings of Workshop on Predictability. 20–22 October 1997. ECMWF, Reading, UK.

Thorarinsdottir, T.L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: Ensemble model output statistics using heteroskedastic censored regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 173, 371–388. https://doi.org/10.1111/j.1467-985X.2009.00616.x

Thorarinsdottir, T.L. and Schuhen, N. (2018). Verification: assessment of calibration and accuracy, In Statistical Post-processing of Ensemble Forecasts, Vannitsem, S., Wilks, D.S., Messner, J.W. (eds), pp. 155–186. Elsevier, Amsterdam, Netherlands.

Vannitsem, S., Wilks, D.S. and Messner, J.W. (Eds.) (2018) *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, Amsterdam, Netherlands.

How to cite this article: Schuhen N, Thorarinsdottir TL, Lenkoski A. Rapid adjustment and post-processing of temperature forecast trajectories. *Q.J.R. Meteorol. Soc.*. 2020;146:963–978. https://doi.org/10.1002/qj.3718

Paper II

Order of operation for multi-stage post-processing of ensemble wind forecast trajectories

Nina Schuhen

In: Nonlinear Processes in Geophysics,
vol. 27, no. 1, pp. 35-49, 2020.
DOI: 10.5194/npg-27-35-2020

Nonlin. Processes Geophys., 27, 35–49, 2020
https://doi.org/10.5194/npg-27-35-2020

Nonlinear Processes
in Geophysics
Open Access

Order of operation for multi-stage post-processing of ensemble wind forecast trajectories

Nina Schuhen

Department for Statistical Analysis, Machine Learning and Image Analysis, Norwegian Computing Center,
P.O. Box 114 Blindern, 0314 Oslo, Norway

Correspondence: Nina Schuhen (nina.schuhen@nr.no)

Received: 24 October 2019 – Discussion started: 30 October 2019
Revised: 5 January 2020 – Accepted: 13 January 2020 – Published: 6 February 2020

Abstract. With numerical weather prediction ensembles unable to produce sufficiently calibrated forecasts, statistical post-processing is needed to correct deterministic and probabilistic biases. Over the past decades, a number of methods addressing this issue have been proposed, with ensemble model output statistics (EMOS) and Bayesian model averaging (BMA) among the most popular. They are able to produce skillful deterministic and probabilistic forecasts for a wide range of applications. These methods are usually applied to the newest model run as soon as it has finished, before the entire forecast trajectory is issued. RAFT (rapid adjustment of forecast trajectories), a recently proposed novel approach, aims to improve these forecasts even further, utilizing the error correlation patterns between lead times. As soon as the first forecasts are verified, we start updating the remainder of the trajectory based on the newly gathered error information. As RAFT works particularly well in conjunction with other post-processing methods like EMOS and techniques designed to reconstruct the multivariate dependency structure like ensemble copula coupling (ECC), we look to identify the optimal combination of these methods. In our study, we apply multi-stage post-processing to wind speed forecasts from the UK Met Office's convective-scale MOGREPS-UK ensemble and analyze results for short-range forecasts at a number of sites in the UK and the Republic of Ireland.

1 Introduction

Numerical weather prediction (NWP) is an inherently uncertain process, and even with present-day computational resources, ensembles can not produce perfect forecasts (Buizza, 2018). Statistical post-processing methods have been successfully applied to address these deficiencies, aiming to resolve a multitude of issues. Two important properties of probabilistic forecasts are calibration and sharpness (Gneiting et al., 2007). Calibration is the statistical consistency between the forecasts and the observations, and sharpness refers to the amount of predictive uncertainty and thus the extent of information contained in the forecast. Usually, NWP ensembles lack calibration, as they can not consider all sources of atmospheric uncertainty, but they are quite sharp. The main goal of any statistical post-processing process should therefore be to maximize the forecast's sharpness, subject to it being calibrated (Gneiting et al., 2007).

Well-established techniques like ensemble model output statistics (EMOS; e.g., Gneiting et al., 2005) or Bayesian model averaging (BMA; e.g., Raftery et al., 2005) are now available for a number of weather variables; for an overview, see Wilks (2018). They measure the ensemble's performance over a training period, either consisting of a rolling window of a few weeks or a longer, fixed period of time, and then apply a statistical correction to the newest NWP model run. The updated forecasts are usually in the form of a predictive probability distribution, as close to perfect calibration as possible. As EMOS has been proven to work well for our data set, the MOGREPS-UK ensemble produced by the UK Met Office, and is computationally more efficient, we prefer it over BMA.

Published by Copernicus Publications on behalf of the European Geosciences Union & the American Geophysical Union.

During the application of some of the methods mentioned above, any physical, spatial and temporal dependency structure from the NWP model is lost and additional effort is needed to restore these patterns (Schefzik and Möller, 2018). In some cases, parametric models can be developed (e.g., Schuhen et al., 2012; Feldmann et al., 2015); however, if this is not feasible, techniques like ensemble copula coupling (ECC; Schefzik et al., 2013) and the Schaake shuffle (Clark et al., 2004) provide a non-parametric approach based on reordering samples from the calibrated predictive distributions. In this study, we choose ECC over the Schaake shuffle, as it does not require any additional historical data.

Recently, Schuhen et al. (2020) proposed a new kind of post-processing method, rapid adjustment of forecast trajectories (RAFT), designed to minimize forecast errors on-the-fly. Instead of running once, like EMOS or BMA, between the NWP model run finishing and the publication or delivery of the forecasts, it is applied repeatedly at every lead time step. RAFT works in concert with conventional post-processing techniques and utilizes the error information from the part of a forecast trajectory where observations are already available in order to improve the mean forecast skill for the rest of the trajectory. This means that, e.g., any systematic forecast error in a model run that was not picked up by the standard post-processing can now be corrected quickly, once it is recorded. In this way, older forecasts become more valuable and typically outperform the first few forecasts of a new model run. While Schuhen et al. (2020) adjust the deterministic mean forecast only, we will show in this paper how RAFT can also be used to adjust the predictive variance. In general, RAFT applies to any kind of forecast scenario, from the short range to seasonal forecasting, as long as there is sufficient correlation between the errors at different lead times.

With an abundance of post-processing methods available, the question arises in which order they should be employed. Li et al. (2019) look at this problem in the context of generator-based post-processing (GPP; Chen and Brissette, 2014), producing discrete, auto-correlated time series, and dependence reconstruction methods like ECC. When working with EMOS, it should generally be run first in order to remove large-scale calibration errors and provide a skillful baseline forecast. However, it is not obvious how to combine ECC and RAFT. Therefore it is our aim to find the optimal order of operation for these three post-processing methods, each designed to achieve a different objective. The combinations of post-processing methods will be applied to site-specific instantaneous wind speed forecasts produced by the high-resolution MOGREPS-UK ensemble and will be assessed using multiple univariate and multivariate verification tools.

The remainder of this paper is organized as follows. In Sect. 2, we introduce the data set used in this study. Section 3 describes the individual post-processing methods, including the new RAFT approach, and Sect. 4 outlines the set of verification metrics which we apply to determine the forecast performance. In Sect. 5, we illustrate how the different techniques work by means of an example forecast and present results for the selected combinations of post-processing methods. We conclude with a discussion in Sect. 6.

2 Data

The 10 m instantaneous wind speed forecast data used in this study were produced by the UK Met Office's limited-area ensemble MOGREPS-UK (Hagelin et al., 2017). MOGREPS-UK is based on the convection-permitting NWP model UKV, but with a lower resolution of 2.2 km. Until March 2016, the global ensemble MOGREPS-G produced both initial and boundary conditions for MOGREPS-UK; subsequently perturbations from the global ensemble were combined with UKV analysis increments to generate the initial conditions.

We use data from all model versions between January 2014 and June 2016, during which the ensemble was initialized four times a day and consisted of 12 members, one control and 11 perturbed forecasts. Here, we only look at the model run started at 15:00 UTC, as it was observed in Schuhen et al. (2020) that all four runs behave somewhat similarly in terms of predictability. Forecasts are produced for every hour up to 36 h, covering the short range.

For both estimation and evaluation, SYNOP observations from 152 sites in the British Isles are used (see Fig. 1). To match the observation locations, the forecasts were interpolated from the model grid and subjected to Met Office post-processing in order to correct for local effects and differences between the model and the location's orography. We separate our data set into two parts: the first 12 months are used for estimating the RAFT coefficients and the remaining 18 months for evaluating the post-processing techniques.

3 Post-processing methods

In this paper, several post-processing methods are used in various combinations. They all fulfill different purposes: EMOS functions as a baseline for producing calibrated and sharp probabilistic forecasts, ECC transfers the physical dependency structure of the ensemble to the EMOS forecasts and RAFT continually improves the EMOS deterministic forecasts after they have been issued, based on previously unavailable information.

3.1 Ensemble model output statistics

In a first step, all forecasts are post-processed with EMOS, sometimes also called non-homogeneous regression, in order to correct deterministic and probabilistic biases the raw ensemble might suffer from. These deficiencies are a result of the limits of ensemble forecasting in general, as, e.g., the ensemble members can only represent a small subset of the multitude of all possible or probable states of the atmo-

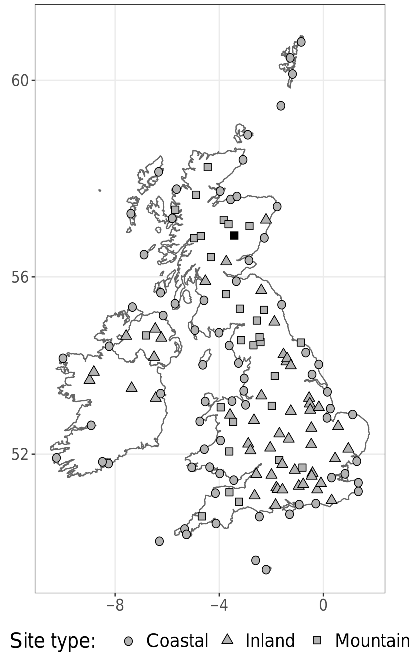


Figure 1. Map of the British Isles with the 152 observation locations used in this study. The sites are divided into three categories, coastal, inland and mountain sites, depending on their location and altitude. The black square marks The Cairnwell, a mountainous site in the Scottish Highlands.

sphere at any given point in time. Thorarinsdottir and Gneiting (2010) propose an application of the EMOS method for wind speed forecasts based on truncated Gaussian distributions, although they study maximum instead of instantaneous wind speed.

As we will see in Sect. 5, this approach (here called gEMOS) produces nearly calibrated forecasts, but they are still slightly underdispersive. For this reason, we investigate a second variant of EMOS introduced by Scheuerer and Möller (2015), logEMOS, where the predictive distributions are truncated logistic. Due to its heavier tails, the logistic distribution can provide a better fit to the instantaneous wind speed data at hand. Further case studies including various versions of EMOS have shown that sharp and calibrated forecasts can be produced for a number of different NWP ensembles (e.g., Feldmann et al., 2015; Scheuerer and Büermann, 2014; Kann et al., 2009).

Let $X_1, \ldots, X_{12}$ denote an ensemble forecast valid at a specific time and location and Y be the corresponding observed wind speed. Then we model the gEMOS forecast as a truncated Gaussian distribution with cut-off at zero, in order to account for the non-negativity of the wind speed values:

$$Y|X_1, \ldots, X_{12} \sim \mathcal{N}^+\left(\mu, \sigma^2\right). \tag{1}$$

Due to the truncation, the negative part of the distribution is cut off and a corresponding probability mass added to the positive part. This means that the parameter μ here is not the mean of the distribution, but the location parameter, and σ^2 is the scale parameter. Using the ensemble mean $\overline{X} = \frac{1}{12}\sum_{i=1}^{12} X_i$ and variance $S^2 = \frac{1}{12}\sum_{i=1}^{12}\left(X_i - \overline{X}\right)^2$ as predictors for the EMOS parameters μ and σ^2, we define the following equations:

$$\mu = a + b^2 \cdot \overline{X}, \tag{2}$$

$$\sigma^2 = c^2 + d^2 \cdot S^2. \tag{3}$$

The coefficients b, c and d are squared in order to simplify interpretability and to make sure that the scale parameter is positive. Minimum score estimation is a versatile way to obtain parameter estimates in such a setting (Dawid et al., 2016). The proper score we want to optimize is the continuous ranked probability score (CRPS; Matheson and Winkler, 1976; Gneiting and Raftery, 2007), which addresses both important forecast properties, sharpness and calibration (for details, see Sect. 4). We process all locations and lead times separately, equivalent to the local EMOS approach in Thorarinsdottir and Gneiting (2010), and the training data consist of a rolling period of 40 d. In practice, this means that the training period contains forecast–observation pairs from the last 40 d preceding the start of the model run, valid at the same lead time and location.

In the case of logEMOS, we substitute the truncated Gaussian distribution in Eq. (1) with a truncated logistic distribution:

$$Y|X_1, \ldots, X_{12} \sim \mathcal{L}^+\left(\mu, s\right), \tag{4}$$

where μ is again the location parameter and $s = \sqrt{3\sigma^2} \cdot \pi^{-1}$ the scale. The location parameter μ and variance σ^2 are linked to the ensemble statistics in the same way as in Eq. (2). Scheuerer and Möller (2015) provide a closed form of the CRPS for a truncated logistic distribution, meaning that gEMOS and logEMOS are comparable in terms of computational cost and complexity. We found parameter estimation to be more stable when applying EMOS to wind speed in knots as compared to meters per second. The ensemble members are treated as exchangeable, in that we use the ensemble mean as a predictor for the EMOS location parameter. This results in more robust parameters and faster computation.

3.2 Ensemble copula coupling

While EMOS is particularly adept at calibrating ensemble forecast, the ensemble's rank structure is lost in the process. To restore the physical dependencies between forecasts at different lead times, we employ ECC (e.g., Schefzik et al., 2013). This method makes use of the original ensemble's multivariate dependency information and transfers it to the new, calibrated forecasts.

First, we draw samples from the univariate EMOS distributions. There are several options, but Schefzik et al. (2013) (as a consequence of the discussion in Bröcker, 2012) recommend using equidistant quantiles, as they best preserve the calibration of the univariate forecasts. Then we reorder the quantiles according to the order statistic of the ensemble members. Thus, for each ensemble member X_i at any given forecast lead time $l = 1, \ldots, 36$, we note its rank among the other ensemble members $X_1^{(l)}, \ldots, X_{12}^{(l)}$. We obtain a permutation τ_l of the numbers $1, \ldots, 12$ such that

$$X_{\tau_l(1)}^{(l)} \leq X_{\tau_l(2)}^{(l)} \leq \ldots \leq X_{\tau_l(12)}^{(l)}. \tag{5}$$

Any ties are resolved at random. Then we apply τ_l to the EMOS quantiles $\widetilde{X}_1^{(l)}, \ldots, \widetilde{X}_{12}^{(l)}$ and reorder the individual ensemble members so that we obtain a multivariate ensemble

$$\left[\widetilde{X}_{\tau_l(1)}^{(1)}, \ldots, \widetilde{X}_{\tau_l(1)}^{(36)}\right], \ldots, \left[\widetilde{X}_{\tau_l(12)}^{(1)}, \ldots, \widetilde{X}_{\tau_l(12)}^{(36)}\right]. \tag{6}$$

The new ensemble has the same univariate properties as the original EMOS quantiles, as only the order of the ensemble members has changed. However, when we evaluate it using multivariate scores and verification tools, we can see the benefit of ECC. It is a computationally efficient and straightforward method to preserve spatial and temporal features of the NWP model. ECC has been used in a variety of atmospheric and hydrological forecasting scenarios, e.g., Schuhen et al. (2012), Hemri et al. (2015) and Ben Bouallègue et al. (2016).

3.3 Rapid adjustment of forecast trajectories

RAFT is a new technique that can be used in conjunction with established approaches like EMOS and ECC. However, it operates on a different timescale. While EMOS and ECC are applied once when the NWP model run has finished, RAFT continually updates the forecast after it has been issued, using information from the part of the forecast trajectory that has already realized. Essentially, RAFT applies to any weather variable; therefore, we do not have to make many alterations to the method for temperature described in Schuhen et al. (2020). We treat all locations separately, as the local error characteristics vary greatly.

In this paper, there are two different RAFT concepts used: we call the standard method that adjusts the EMOS mean RAFT$_\mathrm{m}$, while RAFT$_\mathrm{ens}$ applies to individual ensemble members drawn from the EMOS distribution. RAFT$_\mathrm{m}$ therefore can only improve the deterministic forecast skill, whereas RAFT$_\mathrm{ens}$ provides an adjusted empirical distribution spanned by the ensemble. Both RAFT variants are based on the correlation between observed forecast errors at different lead times. We define the error $e_{t,l}$ at a particular lead time l, generated from a model run started at time t, as the difference between the forecast and the observation y_{t+l}:

$$e_{t,l} = y_{t+l} - m_{t,l}, \tag{7}$$

$$e_{t,l}^{(i)} = y_{t+l} - x_{t,l}^{(i)}, \quad i = 1, \ldots, 12. \tag{8}$$

Equation (7) refers to the RAFT$_\mathrm{m}$ approach, where $m_{t,l}$ is the mean of the EMOS distribution. For RAFT$_\mathrm{ens}$, we need to calculate the error for every ensemble member $x_{t,l}^{(i)}$ (Eq. 8). To obtain the mean of the truncated Gaussian distribution from the location and scale parameters μ and σ^2, we use the following relationship:

$$m = \mu + \sigma \cdot \varphi\left(-\frac{\mu}{\sigma}\right) \cdot \left(1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right)^{-1}. \tag{9}$$

The functions φ and Φ here denote the density and cumulative distribution function of the standard Gaussian distribution, respectively. Similarly, the mean of the truncated logistic distribution is

$$m = s \cdot \log\left(1 + \exp\left(\frac{\mu}{s}\right)\right) \cdot \left(1 - \Lambda\left(-\frac{\mu}{s}\right)\right)^{-1}, \tag{10}$$

where μ is the location parameter, s the scale and Λ the cumulative distribution function (CDF) of the standard logistic distribution.

From the forecast errors $e_{t,l}$ and $e_{t,l}^{(i)}$, we generate the Pearson correlation coefficient matrix to establish the relationship between the 36 lead times. In the RAFT$_\mathrm{ens}$ case this means looking at the correlation matrices of each ensemble member separately. The left column in Fig. 2 shows the gEMOS error correlation matrix for the weather station on The Cairnwell mountain in the Scottish Highlands. The top plot refers to RAFT$_\mathrm{m}$, while the bottom illustrates the correlation for one member of the RAFT$_\mathrm{ens}$ ensemble. All correlations shown are statistically significant at the 90 % level. There is a good correlation between a sizable number of lead times, which makes it possible to define an adjustment period for each lead time, telling us at what point in time to begin with the RAFT adjustments. While the adjustment period applies, we know that a previously observed error e_{t,l^*} at lead time $l^* < l$ gives us reliable information about the future error $e_{t,l}$.

The RAFT model used to obtain the estimated future error $\hat{e}_{t,l}$ at $l = 1, \ldots, 36$ is based on linear regression with the observed error e_{t,l^*} as predictor:

$$\hat{e}_{t,l} = \hat{\alpha} + \hat{\beta} \cdot e_{t,l^*} + \varepsilon, \tag{11}$$

where the random error term ε is normally distributed with mean zero. The regression coefficients $\hat{\alpha}$ and $\hat{\beta}$ are determined using least squares. Once we have estimated the RAFT regression coefficients for every possible combination of lead times l and l^*, we can establish the length p of the individual adjustment periods by looking at those combinations where the estimate of the coefficient $\hat{\beta}$ is significantly greater than zero, meaning that e_{t,l^*} is likely to provide useful

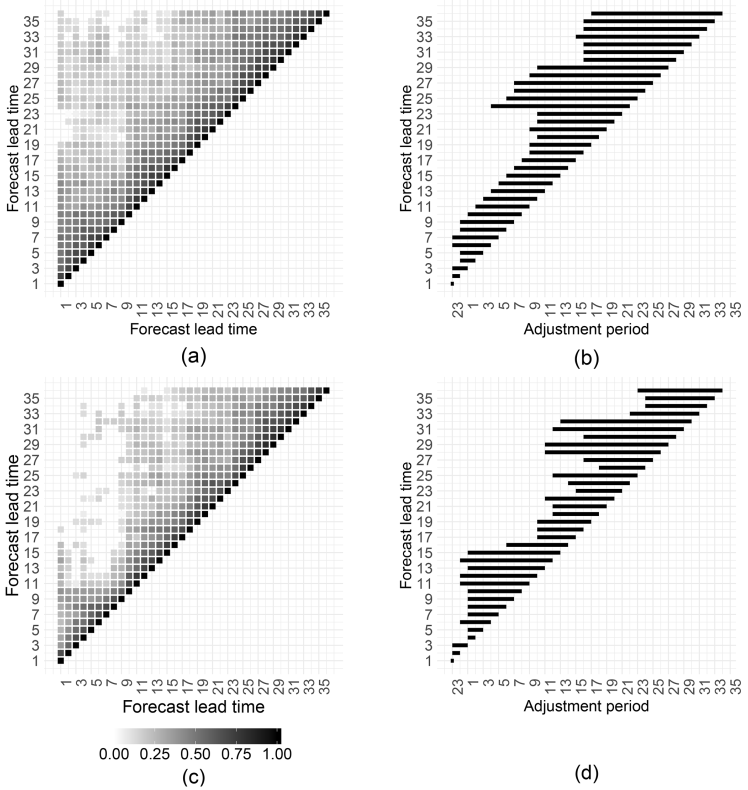


Figure 2. **(a)** Correlation matrix of the EMOS mean error at The Cairnwell. Only correlations significant at the 90 % level are shown. **(b)** Length of the RAFT$_\mathrm{m}$ adjustment period for each lead time. **(c)** As **(a)** but for the error of an ensemble member drawn from the EMOS predictive distribution. **(d)** Length of the RAFT$_\mathrm{ens}$ adjustment period for each lead time.

information for the prediction of $e_{t,l}$. In order to account for potential limitations in real-time availability of observations, the RAFT adjustments performed at a certain lead time $l-1$ for any lead time greater than or equal to l use the observation recorded at $l-2$. For the first few lead times of a model run, where no previous error can be computed, the predictors in Eq. (11) are based on the forecasts from the model run initialized 24 h earlier.

The algorithm for determining the adjustment period corresponds to the one described in Schuhen et al. (2020). In general, it can be applied to any weather variable with errors on a continuous scale. However, it is somewhat arbitrary and can certainly be optimized for individual forecasting scenarios. The algorithm is run once, based on the fixed estimation data set. We proceed as follows.

1. Estimate the regression coefficients in Eq. (11) for all predictors e_{t,l^*} with l^* in $[l-23; l-2]$. If any l^* are negative, we use l^*+24 as predictors instead, so that lead time 23 is followed by lead time $0, 1, 2, \ldots$.

2. a. Find the earliest l^* in $[l-11; l-2]$, such that the coefficient $\hat{\beta}$ is significantly different from zero at the 90 % level for each lead time $l^*+1, \ldots, l-2$. Denote the result as l_p.
   b. If there is no result in the previous step, find the earliest l^* in $[l-19; l-12]$, such that $\hat{\beta}$ is significantly different from zero at the 95 % level for each lead time $l^*+1, \ldots, l-12$. Denote the result as l_p.
   c. If there is no result in the previous step, find the earliest l^* in $[l-23; l-20]$, such that $\hat{\beta}$ is significantly different from zero at the 99 % level for each lead time $l^*+1, \ldots, l-20$. Denote the result as l_p.

3. After running the first two steps for all lead times, determine the length of the adjustment period p.
   a. If Step 2 has yielded a result for l_p, set $p = l - l_p$.

b. If Step 2 has not yielded a result, set p equal to the average of the adjustment period length values for the neighboring lead times $l-1$ and $l+1$.

c. If there is still no valid value for p, set it to $p=22$. This corresponds to the longest possible adjustment period.

Figure 2b and d show the adjustment periods for the RAFT_m (top) and RAFT_ens (bottom) versions at The Cairnwell. For the ensemble method, the algorithm results in a good approximation of the correlation pattern in panel (c), but the values of p seem to jump back and forth with increasing lead time. In the case of the EMOS mean, the values of p are more consistent across the lead times, but do not necessarily correspond as well to the respective correlation matrix pattern in panel (a).

Finding the optimal adjustment periods concludes the estimation part of RAFT. The actual adjustment of the predicted forecast error happens in real time once the current model run has finished and the forecasts have been issued. For lead time l, the adjustment starts at $l-p+1$, using the observation recorded at $l-p$, and then continues hourly until $l-1$. The smaller the gap between l and the time the observation was recorded, the greater the value of the error information and therefore the larger the gain in forecast skill.

In practice, we calculate the observed error according to Eq. (7), plug it into Eq. (11) with the appropriate coefficients $\hat{\alpha}$ and $\hat{\beta}$ and obtain the predictive error $\hat{e}_{t,l}$. Then we can add this forecast to the EMOS mean $m_{t,l}$ for RAFT_m or the ensemble member $x_{t,l}^{(i)}$, $i=1,\ldots,12$ drawn from the EMOS distribution for RAFT_ens:

$$\hat{m}_{t,l} = m_{t,l} + \hat{e}_{t,l}, \tag{12}$$

$$\hat{x}_{t,l}^{(i)} = x_{t,l}^{(i)} + \hat{e}_{t,l}. \tag{13}$$

Any values of $\hat{m}_{t,l}$ and $\hat{x}_{t,l}^{(i)}$ that become negative during this process are set to zero in order to account for the non-negativity of wind speed. While we can use the RAFT-adjusted mean $\hat{m}_{t,l}$ as a deterministic forecast, the corresponding location parameter $\hat{\mu}_{t,l}$ is needed to evaluate the full distribution. For this purpose, we solve Eqs. (9) and (10) numerically for μ. This approach can be quite unstable and has to be done carefully so that the resulting distribution is valid. We then combine the new location parameter with the unchanged EMOS variance and thus obtain a predictive distribution. In the case of RAFT_ens, the ensemble members span a discrete distribution. Therefore, we here adjust not only the deterministic forecast, but also simultaneously the spread of the distribution in an adaptive and flow-dependent way.

4 Evaluation methods

There is a multitude of evaluation methods available to assess both deterministic and probabilistic forecast skill (see, e.g., Thorarinsdottir and Schuhen, 2018). In addition to looking at univariate verification results, we also want to determine the benefit of various combinations of post-processing methods in a multivariate sense.

Proper scoring rules (Gneiting and Raftery, 2007) are useful tools that assign a numerical value to the quality of a forecast and always judge the optimal forecast to have the best score. Usually, they are averaged over a number of forecast cases n. In the deterministic case, the root-mean-square error (RMSE) gives an indication about the forecast accuracy of the mean forecast, be it the mean of a distribution or an ensemble mean. It is defined as

$$\text{RMSE}(F,y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{mean}(F)-y)^2}, \tag{14}$$

where y is the verifying observation corresponding to the predictive distribution F.

To evaluate probabilistic forecasts, the CRPS (Matheson and Winkler, 1976) is an obvious choice. Given the score's robustness, it is often used for parameter estimation, as in the two EMOS variants gEMOS and logEMOS described in Sect. 3.1. A closed form of the CRPS for the truncated Gaussian was derived by Thorarinsdottir and Gneiting (2010) as

$$\begin{aligned}\text{CRPS}\left(\mathcal{N}^+\left(\mu,\sigma^2\right),y\right) &= \sigma\cdot\Phi\left(\frac{\mu}{\sigma}\right)^{-2}\left[\frac{y-\mu}{\sigma}\Phi\left(\frac{\mu}{\sigma}\right)\right.\\ &\left\{2\Phi\left(\frac{y-\mu}{\sigma}\right)+\Phi\left(\frac{\mu}{\sigma}\right)-2\right\} \qquad (15)\\ &\left.+2\varphi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma}\right)-\frac{1}{\sqrt{\pi}}\Phi\left(\sqrt{2}\frac{\mu}{\sigma}\right)\right], \qquad (16)\end{aligned}$$

where Φ is the CDF and φ the PDF (probability density function) of a standard normal distribution. For the truncated logarithmic distribution, a closed form is also available (Scheuerer and Möller, 2015):

$$\begin{aligned}\text{CRPS}\left(\mathcal{L}^+(\mu,s),y\right) &= (y-\mu)\left(\frac{2p_y-1-p_0}{1-p_0}\right) \qquad (17)\\ &+s\left[\log(1-p_0)\right.\\ &-\frac{1+2\log\left(1-p_y\right)+2p_y\text{logit}\left(p_y\right)}{1-p_0}\\ &\left.-\frac{p_0^2\log(p_0)}{(1-p_0)^2}\right]. \qquad (18)\end{aligned}$$

Here, logit$(\cdot)$ is the logit function and $p_0=\Lambda\left(-\mu s^{-1}\right)$ and $p_y=\Lambda\left((y-\mu)s^{-1}\right)$ are values of the CDF of the truncated logistic distribution Λ. To be able to compare all types of forecasts in a fair way, we draw random samples $X_1,\ldots,X_{12}$ from every continuous predictive distribution and evaluate them using the ensemble version of the CRPS:

$$\mathrm{CRPS}_{\mathrm{ens}}\left(X_1,\ldots,X_{12};y\right)=\frac{1}{12}\sum_{i=1}^{12}\left|X_i-y\right|-\frac{1}{2\cdot 12^2}\sum_{i=1}^{12}\sum_{j=1}^{12}\left|X_i-X_j\right|. \qquad (19)$$

Furthermore, we want to assess the level of calibration in a forecast separately, as it is important to prefer the sharpest forecast subject to calibration (Gneiting et al., 2007). To this end, we check the verification rank histogram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand et al., 1997), where we find the ranks of the observation within the forecast ensemble for each forecast case and plot them as a histogram. A ∩-shaped histogram points towards overdispersed forecast distributions, while a ∪ shape means that the forecasts do not exhibit enough spread. For perfect calibration, a flat histogram is a necessary condition, although not sufficient (Hamill, 2001).

The direct equivalent of the CRPS for multivariate forecasts, the energy score (Gneiting and Raftery, 2007), is defined as

$$\mathrm{ES}\left(F,\boldsymbol{y}\right)=\mathbb{E}_F\left\|\boldsymbol{X}-y\right\|-\frac{1}{2}\mathbb{E}_F\mathbb{E}_F\left\|\boldsymbol{X}-\boldsymbol{X}'\right\|, \qquad (20)$$

where $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random vectors drawn from the multivariate distribution F, y is the observation vector and $\|.\|$ is the Euclidean norm. If we replace the absolute value in Eq. (19) with the Euclidean norm, we obtain an analogous version of the energy score for ensemble member vectors. It is also possible to evaluate deterministic forecasts in multiple dimensions using the Euclidean error, which we derive from the energy score by replacing the distribution F with a point measure:

$$\mathrm{EE}\left(F,\boldsymbol{y}\right)=\left\|\boldsymbol{med}_{\mathrm{F}}-\boldsymbol{y}\right\|. \qquad (21)$$

The multivariate point forecast $\boldsymbol{med}_{\mathrm{F}}$ is the spatial median, computed numerically using `R` package `ICSNP` (Nordhausen et al., 2015). It minimizes the sum of the Euclidean distances to the ensemble members.

While the energy score is generally more sensitive to errors in the predictive mean (Pinson and Tastu, 2013), the variogram score proposed by Scheuerer and Hamill (2015) is better at identifying whether the correlation between the components is correct. In addition to following the authors' recommendation and setting the score's order p to 0.5, we assign equal weights to all lead times. The variogram score then becomes

$$\mathrm{VS}\left(F,\boldsymbol{y}\right)=\sum_{i=1}^{36}\sum_{j=1}^{36}\left(\left\|y_i-y_j\right\|^p-\mathbb{E}_F\left\|X_i-X_j\right\|^p\right)^2, \qquad (22)$$

where y_i and y_j are the ith and jth components of the observation vector and X_i and X_j components of a random vector distributed according to F.

Finally, there are several possibilities to check multivariate calibration, like the multivariate rank histogram (Gneiting et al., 2008), the band depth histogram and the average rank histogram (both Thorarinsdottir et al., 2016). We choose to use the latter in this case, as it is less prone to give misleading results than the multivariate rank histogram and more easily interpretable than the band depth histogram. First, a so-called prerank is calculated, corresponding to the average univariate rank of the vector components:

$$\rho_S\left(\boldsymbol{u}\right)=\frac{1}{36}\sum_{i=1}^{36}\mathrm{rank}_S\left(u_i\right), \qquad (23)$$

with $\mathrm{rank}_S\left(u_i\right)$ being the rank of the ith component of a vector $\boldsymbol{u}$ within the combined set of ensemble member and observation vectors $S=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{12},\boldsymbol{y}\}$. Then the multivariate average rank is the rank of the observation prerank in the set $\{\rho_S\left(\boldsymbol{x}_1\right),\ldots,\rho_S\left(\boldsymbol{x}_{12}\right),\rho_S\left(\boldsymbol{y}\right)\}$. The interpretation of the average rank histogram mirrors that of the univariate rank histogram, and errors in the correlation structure present themselves in the same way as dispersion errors in the marginal distributions (Thorarinsdottir and Schuhen, 2018). Visualization of the histograms is taken from Barnes et al. (2019).

5 Results

It is the purpose of this paper to investigate whether there is a preferred order in applying three different kinds of post-processing methods. In particular, it will be important to see whether ECC should be run once, like EMOS, subsequent to the end of the NWP model run, or whether it should be continuously applied every time the RAFT adjustment occurs. Therefore, there are two combinations of methods to be tested: EMOS + RAFT$_{\mathrm{m}}$ + ECC, where RAFT is applied to the EMOS mean only, and EMOS + ECC + RAFT$_{\mathrm{ens}}$, where we adjust the EMOS/ECC ensemble members and thus at the same time the prediction of uncertainty. As we are interested in a comprehensive assessment of the individual combinations' performance, all scores, whether univariate or multivariate, are of equal importance.

5.1 Example forecast

First, we take a look at an example forecast to illustrate how the different RAFT variants work. Figure 3 shows different forecasts issued from the 15:00 UTC model run on 30 October 2015 at The Cairnwell, Scotland. Panels (a) and (c) depict the gEMOS + ECC + RAFT$_{\mathrm{ens}}$ forecasts, where the mean and prediction interval are obtained by the 12 samples drawn from the EMOS distribution. In panels (b) and (d), we have gEMOS + RAFT$_{\mathrm{m}}$ forecasts, with the mean being the RAFT-adjusted mean of the EMOS distribution and the variance the unchanged EMOS predictive variance. Here, we show two different stages in the RAFT adjustment cycle for each combination of post-processing methods. For the top plots, we

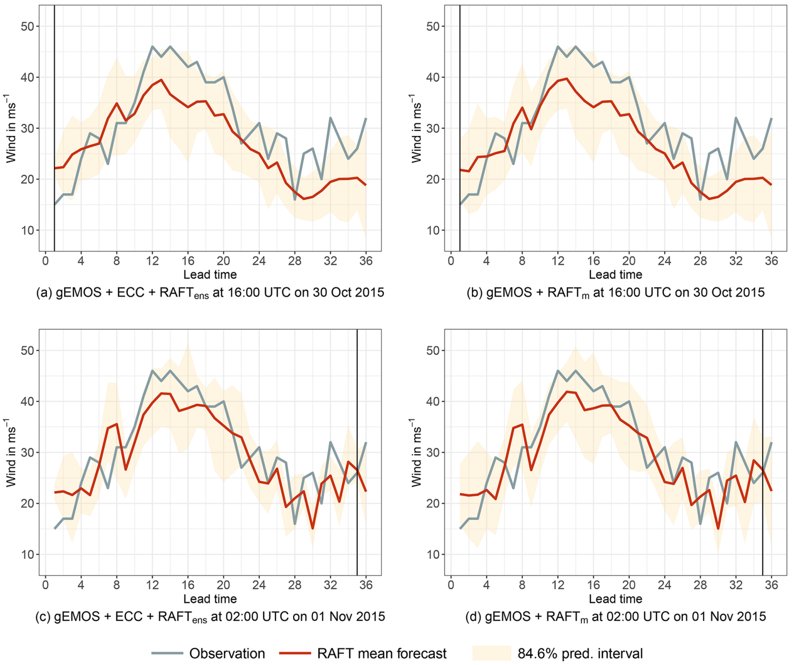


Figure 3. Example forecast at The Cairnwell initialized at 15:00 UTC on 30 October 2015 for the next 36 h. The red line corresponds to the RAFT mean forecast, with the shaded area being the 84.6 % prediction interval. The verifying observation is indicated by the blue line and the vertical line refers to the current point in time during the RAFT adjustment cycle. **(a)** Snapshot of the gEMOS + ECC + RAFT$_{ens}$ forecast taken after RAFT has been applied to the gEMOS + ECC samples once. The prediction interval is spanned by the individually corrected ensemble members. **(b)** Snapshot of the gEMOS + RAFT$_m$ forecast taken after RAFT has been applied to the gEMOS mean once. The prediction interval is based on the gEMOS variance. **(c)** Same as **(a)**, but RAFT has been applied hourly until the last iteration at $t + 35$. **(d)** Same as **(b)**, but RAFT has been applied hourly until the last iteration at $t + 35$.

only apply RAFT once at time $t+1$. This means that all forecasts in the trajectory have been adjusted using the error of the $t+24$ forecast from the model run initialized 24 h earlier, as long as their corresponding adjustment period allows it. The bottom plots are the results of running the whole RAFT adjustment cycle until the last installment at $t+35$. Consequently, all forecasts have been corrected with the observed error measured 2 h earlier and are the most short-term and therefore optimal RAFT forecasts.

In this weather situation, both mean forecasts initially underpredict the wind speed for roughly 12 h starting from lead time 10, corresponding to the time between 01:00 and 13:00 UTC. A further period of underprediction occurs towards the end of the trajectory, from lead time 28. RAFT is able to recognize these problems quickly and corrects the underprediction quite well, as can be observed in the bottom two panels. However, as the observations are quite jumpy, the sign of the forecast error changes frequently during the adjustment process and the RAFT mean trajectory thus can also exhibit more jumpiness than the initial EMOS mean. This could be addressed by, e.g., adding additional predictors to the RAFT linear regression model in Eq. (11).

There are only minor differences in the mean forecasts between the two post-processing method combinations, while their main difference lies in the derivation of the predictive variance. We can see that the size of the prediction interval for gEMOS + ECC + RAFT$_{ens}$ changes considerably between the first and last RAFT adjustments. This is of course because the ensemble, and therefore the prediction interval spanned by its members, is continuously updated and adjusted in a flow-dependent manner. For example, at the end of the trajectory the ensemble spread in Fig. 3c is much smaller than in Fig. 3d. In the case of gEMOS + RAFT$_m$, the variance is not changed by RAFT and remains at the value originally estimated by EMOS.

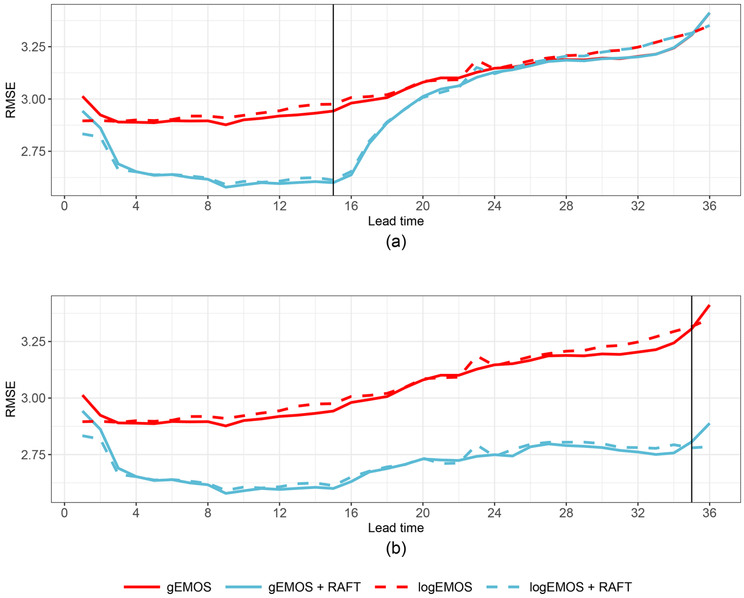


Figure 4. RMSE over lead time for gEMOS (red solid line) and logEMOS (red dashed line) mean forecasts, as well as their RAFT$_m$ adjustments (blue solid and blue dashed lines, respectively). The scores are averaged over all model runs and sites in the evaluation set. **(a)** RAFT is only carried out until the adjustment at $t+15$. **(b)** RAFT is carried out until its last iteration at $t+35$.

Table 1. Univariate and multivariate mean scores for different post-processing method combinations, using the final RAFT adjustments 1 h before valid time. Bold numbers denote the best score in each column. All score differences are significant at the 95 % level, apart from the ones marked with an asterisk, where the pairwise differences between the versions using gEMOS and logEMOS are not significant.

	RMSE	CRPS	Euclidean error	Energy score	Variogram score
Raw ensemble	3.670	2.116	19.207	15.132	847
gEMOS	3.056	1.618	16.539	13.000	956
logEMOS	3.070	1.622	16.589	13.028	957
gEMOS + ECC	3.056	1.618	16.549	12.312	812
logEMOS + ECC	3.070	1.622	16.607	12.356	815
gEMOS + RAFT$_m$	2.713*	1.445	15.045	11.943	899
logEMOS + RAFT$_m$	2.714*	**1.443**	15.029	11.913	897
gEMOS + RAFT$_m$ + ECC	2.713*	1.445	15.049	11.175	**784***
logEMOS + RAFT$_m$ ECC	2.714*	**1.443**	15.033	11.165	**784***
gEMOS + ECC + RAFT$_{ens}$	**2.708***	1.483	15.024*	**11.164***	786
logEMOS + ECC + RAFT$_{ens}$	2.709*	1.482	**15.023***	11.166*	787

5.2 Choice of EMOS model

As we tested two versions of EMOS using two different distributions to model the future wind speed observations, we are interested in which of these, if any, performs better. Initially, we compare the deterministic forecast skill of the EMOS mean and how it is improved by RAFT. In Fig. 4, the RMSE of the gEMOS and logEMOS mean, averaged over all sites and model runs, is shown. Both methods perform very similarly, but gEMOS seems to have a slight advantage overall, apart from the first 3 h and the last hour. There is a small increase in the RMSE for logEMOS at lead time 23, which is most certainly due to issues in finding the minimum CRPS

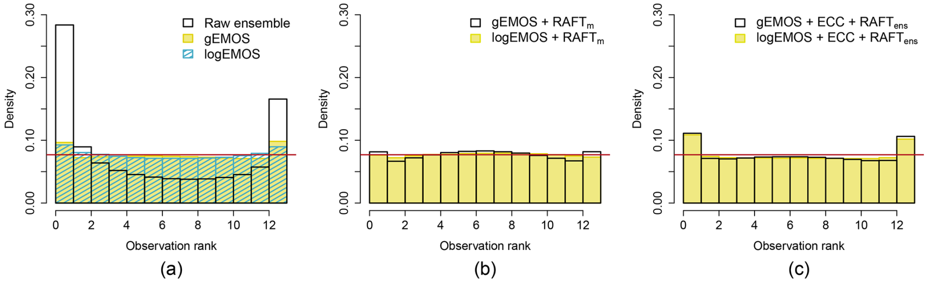


Figure 5. Verification rank histograms for different forecasting methods, aggregated over all sites, model runs and lead times. RAFT histograms are based on the final adjustment for each lead time.

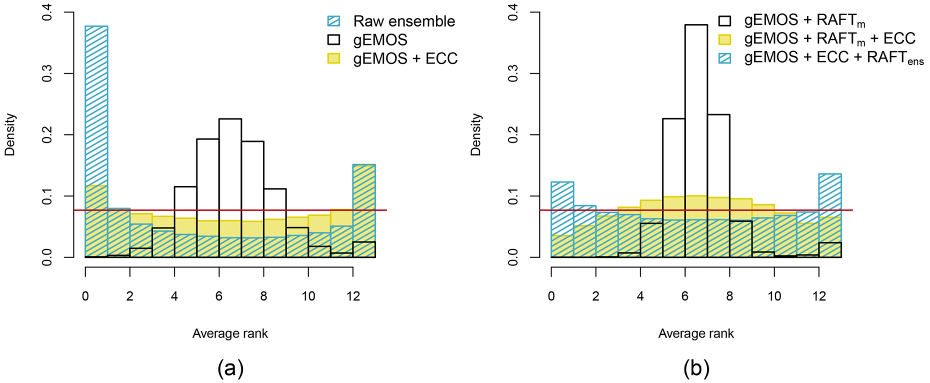


Figure 6. Average rank histogram for different combinations of post-processing methods. Data points are aggregated over all sites, model runs and lead times. All RAFT forecasts have been adjusted using the observation measured 2 h earlier.

during the EMOS parameter estimation, where all lead times are handled separately.

The ranking of the two EMOS variants is preserved when applying RAFT$_m$ to the EMOS mean forecast. Figure 4a shows the RAFT RMSE if we stopped adjusting the forecasts at $t+15$. This means that all forecasts left of the vertical line have been updated using the observation made 2 h earlier, and all forecasts to the right of the line are adjusted using the most recent information available at $t+15$, i.e., the observed error at $t+14$. However, this only applies to those forecasts where lead time 14 lies in the respective adjustment periods. For all other forecasts, the scores for EMOS and RAFT coincide. It is noticeable that the forecast skill improves significantly as soon as we have information about the error in the current model run at $t+3$. The score remains at about the same level until $t+16$, when it starts to deteriorate, but RAFT still has an advantage over the EMOS forecasts for another 10 h. In reality, however, we would run RAFT until the end of the forecast cycle, which is shown in Fig. 4b. Here, we can see a consistent improvement, especially at large lead times. We also see that the forecasts at lead times 25–26 have more skill than the ones at lead times 1–2, which leads to the conclusion that forecasts from a 24 h old model run are for a couple of hours more skillful than the forecast from the newest run.

The first column in Table 1 confirms these results. Here, the scores have been aggregated over all lead times, model runs and sites. In this table only scores for RAFT forecasts that have been adjusted 1 h previously are shown, i.e., the optimal forecasts. We test the significance of score differences by applying a permutation test based on resampling, as described in Heinrich et al. (2019) and Möller et al. (2013). Both EMOS methods increase the deterministic skill considerably when compared to the raw ensemble and then are further improved by applying RAFT$_m$. While the RMSE for gEMOS is significantly better than for logEMOS, which we also see in the CRPS, there is almost no difference in the gEMOS + RAFT$_m$ and logEMOS + RAFT$_m$ scores. In terms of the CRPS, logEMOS + RAFT$_m$ has a slight advantage, with the difference being significant at the 95 % level.

To confirm that the EMOS forecasts are indeed calibrated, we look at the verification rank histograms in Fig. 5a. While

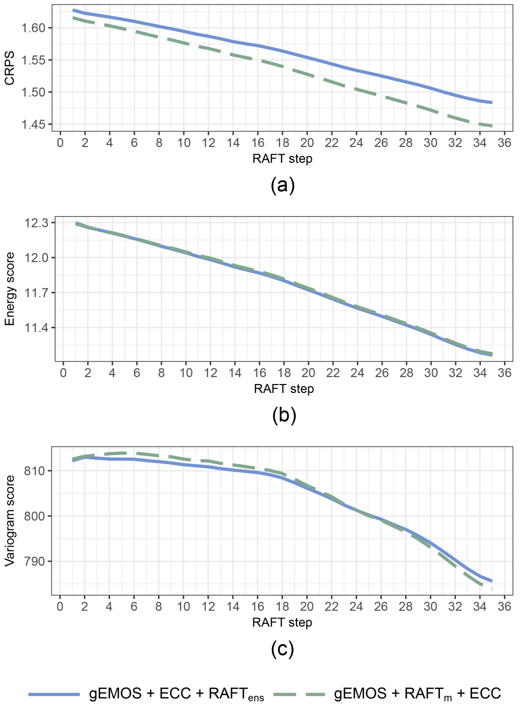


Figure 7. Mean CRPS, energy score and variogram score for every step in the RAFT$_m$ and RAFT$_{ens}$ process. Scores are averaged over all lead times, sites and model runs.

the raw ensemble is very underdispersive, as expected, both gEMOS and logEMOS forecasts are nearly calibrated. Both gEMOS and logEMOS histograms are again very similar, so we compute the coverage of the 84.6 % prediction interval created by 12 ensemble members. From the results we see that logEMOS, with a value of 85.18 %, is much closer to the nominal value than gEMOS with 80.56 % and therefore better calibrated. Figure 5b shows the histograms after we apply RAFT$_m$. Whereas the EMOS variance was on average slightly too small before, it is now a little too big, indicated by the small hump in the middle. This is due to the fact that we do not adjust the variance in this process, but the deterministic skill improves greatly. There is almost no difference in the two histograms, which is also evident in the coverage of the prediction interval, with values of 83.80 % and 83.44 % for gEMOS and logEMOS, respectively.

In conclusion, there is little difference in the overall performance of the two EMOS variants. While logEMOS has the advantage of being slightly better calibrated, gEMOS shows better scores. After applying RAFT, the two methods are essentially equal. In the following we will therefore only present results from one of the two EMOS versions.

5.3 Predictive performance for combinations of post-processing techniques

The main focus of this study is to find out in which order EMOS, RAFT and ECC should be combined. For RAFT, we employ two different approaches: RAFT$_m$, where the adjustments are only applied to the EMOS mean and are then combined with the EMOS variance to obtain a full predictive distribution, and RAFT$_{ens}$, where we adjust individual ensemble members and consequently both mean and spread. In the latter case, ECC is only run once when EMOS has finished; in the first, it has to be applied at every RAFT step for the remaining lead times in the forecast run. Therefore the required computing resources depend on the ratio of ensemble members to lead times. In this study, the EMOS + RAFT$_m$ + ECC combination takes about 33 % more time to compute than EMOS + ECC + RAFT$_{ens}$; however, both are, with only a few seconds per model run and site, computationally very sparse.

When we compare Figure 5b and c, it is obvious that the EMOS + ECC + RAFT$_{ens}$ combination produces forecasts which are slightly less calibrated than EMOS + RAFT$_m$ forecasts. In fact, the level of calibration deteriorates from the baseline EMOS methods. This also can be deduced from the CRPS values in Table 1, where EMOS + RAFT$_m$ clearly performs better. The RMSE for both methods is quite similar, so that we can ascribe the discrepancy in the CRPS to the different levels of calibration. Both methods improve the EMOS baseline forecast considerably. In the case of EMOS + RAFT$_m$, we know this improvement in forecast skill is only due to the adjusted mean forecast, which simultaneously results in better calibrated predictive distributions.

As we are interested not only in the univariate performance, but also in the multidimensional dependencies between the lead times of a forecast trajectory, we look at several multivariate scores (Table 1). The Euclidean error agrees with the univariate RMSE that the best deterministic result can be achieved by applying RAFT last. ECC seems not to have any effect on the scores, which can be expected, as we are only rearranging ensemble members and do not necessarily change the multivariate median. The energy score is a measure for the overall skill, but is also more sensitive to errors in the mean forecast. This is the reason why RAFT$_m$ manages to improve the energy score as compared to EMOS + ECC, while the variogram score deteriorates. Note that both scores are reduced when we reintroduce the temporal correlation structure by applying ECC to the EMOS + RAFT$_m$ forecasts. Although the energy and variogram scores for EMOS + ECC + RAFT$_{ens}$ and EMOS + RAFT$_m$ + ECC are very close, the two scores prefer different post-processing method combinations. While the energy score judges the method to be the best where we apply ECC first, which also has the best RMSE and Euclidean error, the variogram score assigns the lowest value to the better calibrated EMOS + RAFT$_m$ + ECC. The almost identical variogram scores sug-

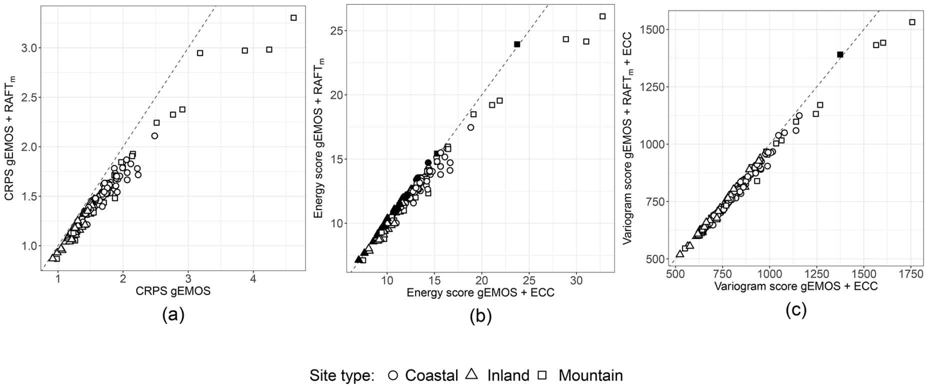


Figure 8. **(a)** CRPS of gEMOS + RAFT$_m$ forecasts against the CRPS of gEMOS forecasts for individual sites. **(b)** Energy scores of gEMOS + RAFT$_m$ forecasts against energy scores of gEMOS + ECC forecasts for individual sites. **(c)** Variogram scores of gEMOS + RAFT$_m$ + ECC forecasts against variogram scores of gEMOS + ECC forecasts for individual sites. The scores are averaged over all lead times and model runs. RAFT forecasts are taken from the final iteration. Filled symbols denote that the score on the y axis is lower than the one on the x axis; empty symbols denote the opposite.

gest that RAFT$_{ens}$ manages to preserve the multivariate correlation structure throughout its multiple iterations.

The average rank histograms in Fig. 6 confirm that without applying ECC, the EMOS and EMOS + RAFT$_m$ forecasts are very uncalibrated. Both the EMOS + ECC and EMOS + ECC + RAFT$_{ens}$ combinations show a ∪-like shape, which can be interpreted as either a too strong correlation or underdispersion. From the band depth histogram (not shown; see Thorarinsdottir et al., 2016) we can conclude that the latter is the case here, as was also seen in the univariate histograms. On the other hand, the EMOS + RAFT$_m$ + ECC forecast ranks form a hump-like histogram. This is due to the correlation between the components being too weak, again confirmed by the band depth histogram.

In order to investigate further the optimal order of operation when applying multiple post-processing methods, we look at how the scores develop with every step in the RAFT process. While the scores in Table 1 are computed using the final RAFT installment at $t+35$, where all forecasts have been adjusted using the observation made 2 h earlier, Fig. 7 shows the CRPS, energy score and variogram score computed at each RAFT iteration for the gEMOS + ECC + RAFT$_{ens}$ and gEMOS + RAFT$_m$ + ECC forecasts. From Fig. 7a, it is clear that EMOS + RAFT$_m$ + ECC performs best in terms of the CRPS, with the gap between the two combinations widening with increasing number of RAFT adjustments. As we have also seen that the RAFT$_m$ version is better calibrated than the RAFT$_{ens}$ one, this means that the CRPS here puts more weight on the calibration being correct than on the slightly better deterministic forecast (see the RMSE in Table 1) in the RAFT$_{ens}$ case. This is surprising, given that the CRPS and its multivariate equivalent, the energy score, are usually more sensitive to the error in the forecast mean (see Fig. 4 in Friederichs and Thorarinsdottir, 2012, and Pinson and Tastu, 2013).

While the CRPS results show a clear pattern, it is not as straightforward for the energy score. The mean score decreases with every RAFT adjustment, as expected, but there is no discernible difference in the performance of the two post-processing method combinations. The most complex picture emerges in the case of the variogram score, where the ranking of the two combinations actually switches around RAFT iteration 24. The variogram score is better at detecting incorrect correlation structures than the energy score, so one possible explanation would be that EMOS + ECC + RAFT$_{ens}$ is initially good at retaining the appropriate correlations, but that ability weakens over time. Conversely, ECC is applied after every iteration of RAFT$_m$, which might explain the better variogram scores towards the end of the process. However, we have observed in Fig. 6 that the correlation structure at the last iteration is still too weak. It should also be noted that the variogram score for EMOS + RAFT$_m$ + ECC deteriorates slightly at the beginning.

Finally, we want to investigate the homogeneity of the scores across the different locations and highlight some interesting results for particular sites. In Fig. 8a, we see that RAFT$_m$ improves the CRPS for all sites as compared to the EMOS baseline. That means that the method where the variance is not adjusted increases the deterministic and probabilistic forecast skill at all sites. As we have seen from the univariate histograms in Fig. 5, even the calibration is improved. Figure 8b shows how a reduction in the mean error can have a large effect on the energy score. At 37 sites, the energy score for gEMOS + RAFT$_m$ is actually lower than the one for gEMOS + ECC. The former forecasts are lacking any form of temporal coherency among lead times, so here the deterministic improvement exceeds any benefit from reintroducing the ensemble's correlation information. Judging from Fig. 8c, a case can be made for a site-specific RAFT approach. The mean variogram score for the gEMOS + RAFT_m + ECC forecasts at The Cairnwell, Scotland, is higher than the score for gEMOS + ECC, meaning that we are not able to make any improvements by applying RAFT_m and that there are local effects not resolved by the RAFT model.

6 Conclusions

Our goal was to find out in which order post-processing methods pertaining to different stages in the forecasting process should be applied. We look at three techniques, each with a different objective. EMOS is a versatile method aiming to calibrate ensemble output as soon as the model run is finished, based on the ensemble's performance over the last 40 d. There are two candidates for wind speed calibration: gEMOS uses a model based on truncated Gaussian distributions and logEMOS a model based on truncated logarithmic distributions. It turns out that both models produce very similar results, with gEMOS having slightly better scores and logEMOS being a little closer to perfect calibration. Therefore it is advisable to test both methods for the data set at hand and to check which distribution gives a better fit.

The second technique, ECC, restores the multivariate dependency structure present in the ensemble forecasts to the EMOS predictions. While conceptually and computationally easy to implement, the success of ECC relies on the NWP model getting the physical, spatial and temporal correlations between the components right. Making use of the part of a forecast trajectory that has already been verified, RAFT is based on the concept that an observed error will provide information about the expected error at not-yet-realized lead times. It can be applied either to the forecast mean only (RAFT_m) or to a set of ensemble members (RAFT_ens) in order to adjust both predictive mean and variance.

In essence, there are two feasible options when combining these three methods: EMOS + ECC + RAFT_ens and EMOS + RAFT_m + ECC. Overall, their performance might be very similar, but there are subtle differences which can lead to preferring one method over the other. The EMOS + RAFT_m + ECC variant produces a lower CRPS and has better univariate calibration, although this is most likely a feature of this forecasting system only, where the EMOS forecasts are underdispersive. Naturally, the RAFT_ens adjusted predictive variance becomes smaller with every RAFT step, as predictability usually increases with a shrinking forecast horizon. This, however, leads to the respective distributions still being underdispersed and not able to counterbalance the deficit of the EMOS forecasts.

If multivariate coherency is of particular importance, e.g., to create plausible forecast scenarios, the EMOS + ECC + RAFT_ens turns out to be the better choice, as is beats its alternative in terms of the energy score, the Euclidean error and also the RMSE, while there is only very little difference in the variogram score. It is also more versatile and should be preferred for NWP ensembles exhibiting very different calibration characteristics than MOGREPS-UK.

Therefore, it is necessary to study every forecasting scenario closely, monitor how calibration methods like EMOS affect the forecast skill and identify potentially remaining deficiencies. As a rule of thumb, it can be said that the post-processing method designed to address one's particular area of interest, whether univariate or multivariate, should be applied first. So far, we do not adapt RAFT to optimize forecasts at individual sites. A model tailored to specific local characteristics could involve changing the algorithm for finding the adjustment period or adding suitable predictors to the linear model. Also, particular attention should be paid to whether the focus lies on a specific subset of lead times or whether the forecasts have to be irrevocably issued at a certain point in time, as the ranking of methods can change during the RAFT process.

Data availability. The data used in this study are available from the Met Office, FitzRoy Road, Exeter, UK. Data can be requested from the author and are only shared subject to permission from the Met Office.

Author contributions. NS designed and executed the study and composed the manuscript.

Competing interests. The author declares that there is no conflict of interest.

Special issue statement. This article is part of the special issue "Advances in post-processing and blending of deterministic and ensemble forecasts". It is not associated with a conference.

Acknowledgements. The author would like to thank the IMPROVER team at the UK Met Office for the opportunity to work with MOGREPS-UK data and Thordis L. Thorarinsdottir for numerous helpful discussions and feedback.

Financial support. This research has been supported by the Norwegian Research Council (grant no. 259864).

Review statement. This paper was edited by Stephan Hemri and reviewed by two anonymous referees.

References

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:amfpae>2.0.co;2, 1996.

Barnes, C., Brierley, C. M., and Chandler, R. E.: New approaches to postprocessing of multi-model ensemble forecasts, Q. J. Roy. Meteor. Soc., 145, 3479–3498, https://doi.org/10.1002/qj.3632, 2019.

Ben Bouallègue, Z., Heppelmann, T., Theis, S. E., and Pinson, P.: Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach, Mon. Weather Rev., 144, 4737–4750, https://doi.org/10.1175/mwr-d-15-0403.1, 2016.

Bröcker, J.: Evaluating raw ensembles with the continuous ranked probabilty score, Q. J. Roy. Meteor. Soc., 138, 1611–1617, https://doi.org/10.1002/qj.1891, 2012.

Buizza, R.: Ensemble forecasting and the need for calibration, in: Statistical postprocessing of ensemble forecasts, edited by: Vannitsem, S., Wilks, D. S., and Messner, J. W., chap. 2, 15–48, Elsevier, Amsterdam, Netherlands, https://doi.org/10.1016/B978-0-12-812372-0.00002-9, 2018.

Chen, J. and Brissette, F. P.: Postprocessing of ensemble weather forecats using a stochastic weather generator, Mon. Weather Rev., 142, 1106–1124, https://doi.org/10.1175/MWR-D-13-00180.1, 2014.

Clark, M. P., Gangopadhyay, S., Hay, L. E., Rajagopalan, B., and Wilby, R. L.: The Schaake Shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields, J. Hydrometeorol., 5, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2, 2004.

Dawid, A. P., Musio, M., and Ventura, L.: Minimum scoring rule inference, Scand. J. Stat., 43, 123–138, https://doi.org/10.1111/sjos.12168, 2016.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L.: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression, Mon. Weather Rev., 143, 955–971, https://doi.org/10.1175/mwr-d-14-00210.1, 2015.

Friederichs, P. and Thorarinsdottir, T. L.: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction, Environmetrics, 23, 579–594, https://doi.org/10.1002/env.2176, 2012.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T., Raftery, A., Westveld, A., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133, 1098–1118, https://doi.org/10.1175/mwr2904.1, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. Roy. Stat. Soc. B, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A.: Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder), Test, 17, 211–264, https://doi.org/10.1007/s11749-008-0114-x, 2008.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N., and Tennant, W.: The Met Office convective-scale ensemble, MOGREPS-UK, Q. J. Roy. Meteor. Soc., 143, 2846–2861, https://doi.org/10.1002/qj.3135, 2017.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:iorhfv>2.0.co;2, 2001.

Hamill, T. M. and Colucci, S. J.: Verification of Eta-RSM short-range ensemble forecasts, Mon. Weather Rev., 125, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:voersr>2.0.co;2, 1997.

Heinrich, C., Hellton, K. H., Lenkoski, A., and Thorarinsdottir, T. L.: Multivariate postprocessing methods for high-dimensional seasonal weather forecasts, arXiv:1907.09716v2, 2019.

Hemri, S., Lisniak, D., and Klein, B.: Multivariate postprocessing techniques for probabilistic hydrological forecasting, Water Resour. Res., 51, 7436–7451, https://doi.org/10.1002/2014wr016473, 2015.

Kann, A., Wittmann, C., Wang, Y., and Ma, X.: Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis, Mon. Weather Rev., 137, 3373–3387, https://doi.org/10.1175/2009mwr2793.1, 2009.

Li, X., Chen, J., Xu, C., Zhang, X. J., Guo, Q., and Xiong, L.: Postprocessing ensemble weather forecasts for introducing multisite and multivariate correlation using rank shuffle and copula theory, in review, 2019.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Manage. Sci., 22, 1087–1096, https://doi.org/10.1287/mnsc.22.10.1087, 1976.

Möller, A., Lenkoski, A., and Thorarinsdottir, T. L.: Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas, Q. J. Roy. Meteor. Soc., 139, 982–991, https://doi.org/10.1002/qj.2009, 2013.

Nordhausen, K., Sirkia, S., Oja, H., and Tyler, D. E.: ICSNP: Tools for multivariate nonparametrics, R package version 1.1-0, 2015.

Pinson, P. and Tastu, J.: Discrimination ability of the energy score, Tech. rep., Technical University of Denmark (DTU), 2013.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, https://doi.org/10.1175/MWR2906.1, 2005.

Schefzik, R. and Möller, A.: Ensemble postprocessing methods incorporating dependence structures, in: Statistical postprocessing of ensemble forecasts, edited by: Vannitsem, S., Wilks, D. S., and Messner, J. W., chap. 4, 91–125, Elsevier, Amsterdam, Netherlands, https://doi.org/10.1016/B978-0-12-812372-0.00004-2, 2018.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T.: Uncertainty quantification in complex simulation models using ensemble copula coupling, Stat. Sci., 28, 616–640, https://doi.org/10.1214/13-STS443, 2013.

Scheuerer, M. and Büermann, L.: Spatially adaptive post-processing of ensemble forecasts for temperature, J. Roy. Stat. Soc. C, 63, 405–422, https://doi.org/10.1111/rssc.12040, 2014.

Scheuerer, M. and Hamill, T. M.: Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities, Mon. Weather Rev., 143, 1321–1334, https://doi.org/10.1175/mwr-d-14-00269.1, 2015.

Scheuerer, M. and Möller, D.: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics, Ann. Appl. Stat., 9, 1328–1349, https://doi.org/10.1214/15-aoas843, 2015.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T.: Ensemble model output statistics for wind vectors, Mon. Weather Rev., 140, 3204–3219, https://doi.org/10.1175/mwr-d-12-00028.1, 2012.

Schuhen, N, Thorarinsdottir, TL, Lenkoski, A. Rapid adjustment and post-processing of temperature forecast trajectories, Q. J. Roy. Meteor. Soc., 1–16, https://doi.org/10.1002/qj.3718, 2020.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Proc. Workshop on Predictability, 1–25, Reading, UK, European Centre for Medium-Range Weather Forecasts, available at: https://www.ecmwf.int/node/12555 (last access: 31 January 2020), 1997.

Thorarinsdottir, T. L. and Gneiting, T.: Probabilistic forecasts of wind speed: ensemble model output statistics using heteroskedastic censored regression, J. Roy. Stat. Soc. A, 173, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x, 2010.

Thorarinsdottir, T. L. and Schuhen, N.: Verification: assessment of calibration and accuracy, in: Statistical postprocessing of ensemble forecasts, edited by: Vannitsem, S., Wilks, D. S., and Messner, J. W., chap. 6, 155–186, Elsevier, Amsterdam, Netherlands, https://doi.org/10.1016/b978-0-12-812372-0.00006-6, 2018.

Thorarinsdottir, T. L., Scheuerer, M., and Heinz, C.: Assessing the calibration of high-dimensional ensemble forecasts using rank histograms, J. Comput. Graph. Stat., 25, 105–122, https://doi.org/10.1080/10618600.2014.977447, 2016.

Wilks, D.: Univariate ensemble postprocessing, in: Statistical postprocessing of ensemble forecasts, edited by: Vannitsem, S., Wilks, D. S., and Messner, J. W., chap. 3, 49–89, Elsevier, Amsterdam, Netherlands, https://doi.org/10.1016/B978-0-12-812372-0.00003-0, 2018.

Paper III

Trajectory adjustment of lagged seasonal forecast ensembles

Thordis L. Thorarinsdottir, Nina Schuhen, Alex Lenkoski

Technical report 19/20. Norsk Regnesentral. 2020.

III

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Trajectory adjustment of lagged seasonal forecast ensembles

Note no SAMBA/19/20

Authors Thordis L. Thorarinsdottir
Nina Schuhen
Alex Lenkoski

Date 10th June 2020

The authors

Thordis L. Thorarinsdottir and Alex Lenkoski are Chief Research Scientists at the Norwegian Computing Center, Nina Schuhen is Researcher at CICERO Center for International Climate Research

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Title	**Trajectory adjustment of lagged seasonal forecast ensembles**
Authors	**Thordis L. Thorarinsdottir , Nina Schuhen , Alex Lenkoski**
Date	10th June 2020
Publication number	SAMBA/19/20

Abstract

Seasonal forecasting has became a critical area of development in numerical weather prediction. Reliable forecasts beyond the two week time period are necessary for a number of industrial and societal planning applications and new approaches are being developed to extend the useful range of numerical weather prediction output. We investigate the performance of one such system, the UK Met Office's GloSea5 system, an ensemble system with the novel feature that ensemble members are initiated in a rolling and staggered manner. Focusing on summer surface temperatures, we show that individual model runs from this system do not exhibit skill beyond the two-week time horizon and indeed substantially under-perform climatological forecasts at longer lead times. However, when combining the ensemble system and applying the Rapid Adjustment of Forecast Trajectories (RAFT) methodology to the individual runs, we show that the combined forecast can achieve performance which is always at least on par with climatology and in many circumstances exhibits modest outperformance.

Keywords	Seasonal forecasting; Lagged ensemble; Postprocessing; Rapid adjustment of forecast trajectories (RAFT)
Target group	Scientists
Availability	Open
Project	PostProcessingPhD
Project number	220783
Research field	Statistics; Meteorology
Number of pages	13
© Copyright	Norwegian Computing Center og forfatterne

1 Introduction

Weather forecasting beyond the medium range of two weeks is currently an active area of research (Robertson and Vitart, 2018) due to the demand for skillful long-range forecasts in various societal sectors such as energy production, agriculture, health and disaster management (e.g. Ogallo et al., 2008). Sources of long-range predictability within the atmosphere are usually associated with the existence of different modes of low-frequency variability, including the El Niño Southern Oscillation (ENSO), monsoon rains, sudden stratospheric warmings, the Madden Julian Oscillation (MJO), the Indian Ocean dipole, the North Atlantic Oscillation (NAO), and the Pacific/North American (PNA) pattern, spanning a wide range of time scales from months to decades (Hoskins, 2013; Vitart et al., 2012). It is expected that, if a forecasting system is capable of reproducing phenomena with low-frequency variability, they may also be able to forecast them (Van Schaeybroeck and Vannitsem, 2018). Post-processing and skill assessment of long-range forecasts is thus often focused on these same phenomena (e.g. Van Schaeybroeck and Vannitsem, 2018), or other slowly-evolving components of the Earth system such as sea-surface temperature (e.g. Heinrich et al., 2019). However, forecast users commonly need information on atmospheric variables such as surface temperature and precipitation (Roulin and Vannitsem, 2019).

At time scales beyond the medium range, the weather noise that arises from the growth of the initial uncertainty, becomes large (Royer, 1993). As a consequence, predictions must be probabilistic in nature. This is made possible through the use of ensemble forecasts (Van Schaeybroeck and Vannitsem, 2018), with a trade-off between increased computational costs and increased skill as the ensemble size grows. For monthly to seasonal forecasts, the benefit of good initialization (initialization as close as possible to observations) has been demonstrated (Doblas-Reyes et al., 2013a,b). For these reasons, the UK Met Office's seasonal prediction system, GloSea5, uses a lagged initialization approach with new ensemble members initialized every day, resulting in a monthly seasonal forecast ensemble with 42 members generated by combining all forecasts available from the most recent three weeks (MacLachlan et al., 2015)[1].

In this paper, we investigate how the older members of a lagged ensemble system can be brought closer to observations by utilizing new observations that have become available since the forecast system was run to generate these members, using the rapid adjustment of forecast trajectories (RAFT) algorithm recently proposed by Schuhen et al. (2020). With a focus on weekly average surface temperature, we aim to assess the skill of the forecast in a user-relevant setting. For observations, we use the ERA5 reanalysis. Preliminary data for ERA5 is now being released daily with a 5-day delay from real time, making the setting considered here somewhat realistic from an operational perspective. The data sets and the RAFT algorithm are described in the following Section 2, with results shown in Section 3. Finally, some concluding remarks are given in Section 4.

[1] https://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/user-guide/technical-glosea5

2 Data and methods

2.1 Data

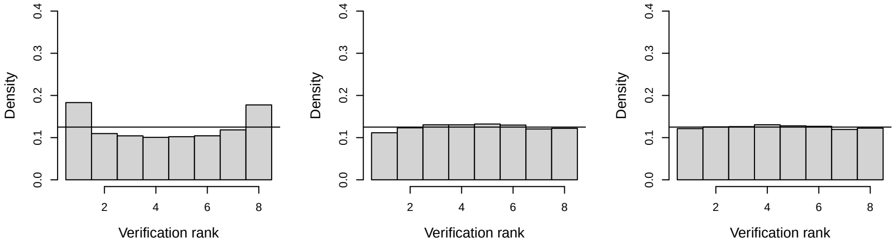


Figure 1. Verification rank histograms for GloSea5 forecasts of weekly mean temperature anomalies initialized on May 1st compared against ERA5. The results are aggregated over the study region, the time period 1993-2015 as well as lead times 1-6 weeks (left), 7-12 weeks (middle) or 13-18 weeks (right). The black horizontal lines indicate a perfectly calibrated forecast.

We analyze surface temperature hindcasts, or historical re-forecasts, from GloSea5, the UK Met Office Global Seasonal forecast system version 5 (MacLachlan et al., 2015). The GloSea5 system has a spatial resolution of 0.8 degrees in latitude and 0.5 degrees in longitude. Our analysis focuses on land grid cells in a region bounded by -30 to 50 longitude and 30 to 90 latitude, covering Europe and surrounding area. The hindcasts cover the time period 1993 to 2015, and the system uses a lagged initialization approach with seven members initialized on the 1st, 9th, 17th and 25th of every month. Hindcasts of weekly mean temperatures from five initialization dates–May 1st to June 1st–are considered for realization dates of up to 18 weeks ahead for the May 1st run, or the time period from early May to early September. The analysis is performed on temperature anomalies which are defined relative to the model's weekly climatology over the entire time period 1993-2015. In the remainder of the paper, we will refer to the hindcasts as "forecasts".

The GloSea5 forecasts are compared against the ERA5 reanalysis (Copernicus Climate Change Service (C3S), 2017). ERA5 originally has a spatial resolution of 0.28 degrees and is here upscaled to match the resolution of the GloSea5 system. We calculate weekly mean anomalies in the same manner as for the hindcasts using ERA5's climatology over the same time period.

The aim of the forecast system is to provide accurate and calibrated forecasts (e.g. Thorarinsdottir and Schuhen, 2018). Calibration, or reliability, refers to the representation of uncertainty in the forecast in that an event predicted to occur with probability p should be realized with the same frequency in the reanalysis. An empirical calibration assessment of the seven member ensemble initialized on May 1st is shown in Figure 1. The plots show the distribution of the rank of the reanalysis when compared against the seven ensemble members across years, spatial locations and forecast lead times. While the forecasts are slightly underdispersive for the first six weeks as indicated by the ∪-shape, they are nearly perfectly calibrated for weeks 7-18. For this reason, we will in the following focus on improving the prediction accuracy.

2.2 Rapid adjustment of forecast trajectories (RAFT)

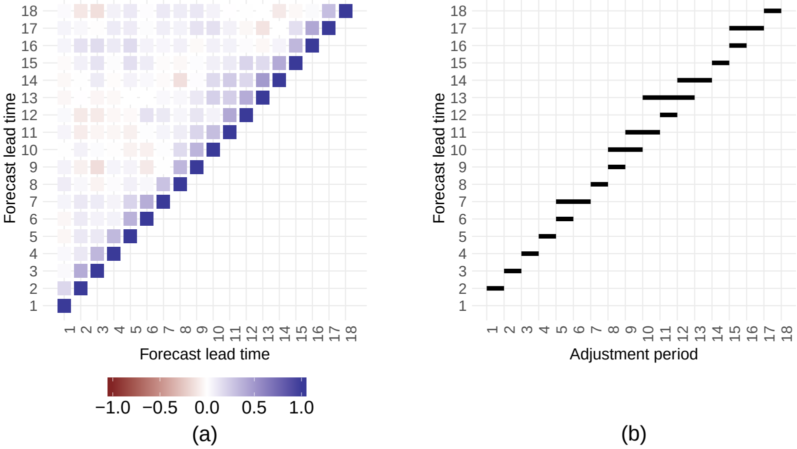


Figure 2. (a) Correlations between forecast anomaly errors at different lead times of the same forecast trajectory for the ensemble mean forecast initialized on May 1st; (b) The resulting adjustment periods for each forecast lead time.

To improve the accuracy of the forecasts, we consider new information that has become available since the forecast was issued, namely observations associated with lead times that have already been realized. Specifically, if the forecast errors at subsequent lead times are correlated with the most recently observed forecast error, this information can be used to update the remaining forecast trajectory that is yet to be realized using the rapid adjustment of forecast trajectories (RAFT) algorithm proposed by Schuhen et al. (2020). The forecast error $e_{t,l}$ is here defined as the distance of the ensemble mean anomaly forecast $\bar{x}_{t,l}$ initialized at time t and valid at lead time l to the observed anomaly y_{t+l} at time $t+l$,

$$e_{t,l} = y_{t+l} - \bar{x}_{t,l}. \tag{1}$$

Figure 2(a) shows the correlation between forecast errors at different lead times for the ensemble mean forecast trajectory initialized on May 1st. While the errors at all lead times beyond the first show substantial correlation with the error observed at the previous lead time, the correlation decreases rapidly for lead times further into the future.

We use a linear regression model to connect the error at a future lead time $l' > l$ with the current error $e_{t,l}$. Specifically, we define the model

$$e_{t,l'} = \alpha + \beta\, e_{t,l} + \varepsilon, \tag{2}$$

where α and β are real valued regression coefficients and ε is a normally distributed error term with mean zero. The model is estimated separately for each forecast run, current lead time and future lead time in a leave-one-out cross-validation approach, i.e. forecast anomaly errors for each year are predicted by using data from all remaining years. In

Schuhen et al. (2020) and Schuhen (2019), the number of future lead times that are corrected each time is selected based on a hypothesis test for $\beta = 0$ after estimating the regression equation in (2) for future lead times $l+1, l+2, \ldots$. At the first future lead time l^* where this test is not rejected, the procedure is stopped and only lead times l' with $l < l' < l^*$ are corrected. Here, this approach turns out to produce unrealistically long adjustment periods and thus spurious correlations can result in reduction of the forecast accuracy rather than an improvement. As we only have a small number of lead times, we instead determine the length of the adjustment periods empirically.

For each l' with $l < l' < l^*$, we then update the ensemble mean forecast $\bar{x}_{t,l'}$ to $\bar{x}_{t,l'} + \hat{e}_{t,l'}$ where $\hat{e}_{t,l'}$ is the estimated error based on (2). The adjustment periods for the forecast run initialized on May 1st are shown in Figure 2(b). Further details of the RAFT algorithm are given in Schuhen et al. (2020) and Schuhen (2019).

3 Results

For evaluating the forecasts, we calculate the root mean square error (RMSE) skill score of the ensemble mean forecast with the ERA5 climatology forecast of that week and grid cell over the entire time period 1993-2015 as a reference forecast. A positive skill score indicates a higher skill than the climatology, while a negative skill score indicates a lower skill.

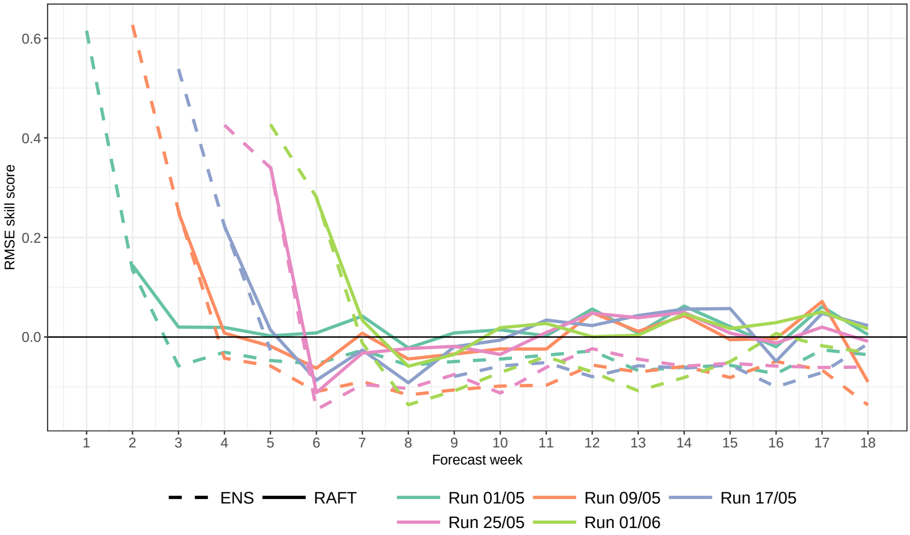


Figure 3. Root mean squared error (RMSE) skill scores for five different runs of GloSea5 compared against ERA5 climatology. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015. The original GloSea5 forecasts are indicated with dashed lines while RAFT forecasts updated one week prior to the realization time are indicated by solid lines.

Figure 3 shows the RMSE skill scores for each of the GloSea5 runs as a function of lead time, aggregated over grid cell locations and years. All the runs show a similar pattern: In the first week, the forecast improves the climatological reference forecast by 40-60%, and in the second week, the forecasts are 15-35% better than climatology. From week three and onward, however, the skill is roughly constant at 5-15% below climatology. At the shortest possible adjustment lead time of one week, the forecasts updated with the RAFT algorithm are consistently better than the original forecasts and, on average, more skillful than the climatology forecast. The skill of RAFT forecasts with adjustment lead times from one week to that of the original forecast generally falls between the two forecasts shown in Figure 3. For example, for the run initialized on May 1st, Figure 2(b) shows that the adjustment period for this run varies from one to three weeks depending on the week. At any given time, the RAFT forecast trajectory will thus converge to the original forecast trajectory after one to three weeks. Results for the other four runs are similar (results not shown). On their own, the individual runs thus do not provide forecasts of higher skill than climatology beyond the medium range of two weeks.

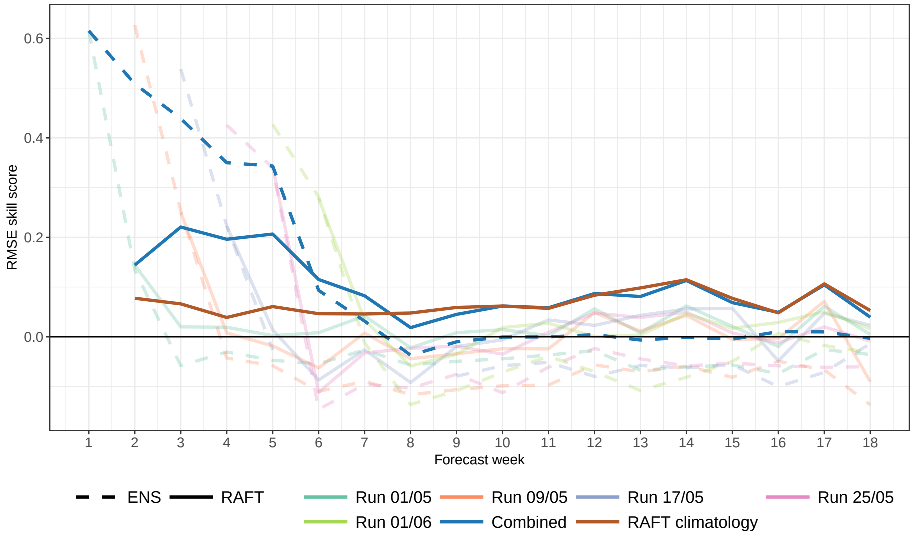


Figure 4. Root mean squared error (RMSE) skill scores for a comparison against ERA5 climatology for the combination of all five forecast runs (blue dashed line), for RAFT-processed ERA5 climatology of adjustment lead time one week (brown solid line), and for a combination of the RAFT-processed ensemble means for all five forecast runs at an adjustment lead time of one week (blue solid line). For comparison, these skill scores are overlaid on the results shown in Figure 3.

We now consider various forecast combinations where, in each case, the multi-model or lagged ensemble mean forecast is constructed using equal weights on the different models. As shown in Figure 4, for the first five forecast weeks, the skill of the lagged ensemble mean is slightly below that of the newest forecast run. From forecast week six and onward, no new runs are added to the lagged ensemble mean, resulting in increasing effective lead time of the forecast. While this gradually reduces the skill, as expected, the reduction halts at around the skill of the climatology and beyond week eight, the two forecasts are comparable in skill. Thus, while the skill of each individual run is lower than that of climatology beyond the medium range of two weeks, their joint skill is consistently higher for lead times of up to three weeks and comparable thereafter.

The climatological reference forecast may be updated in the same manner as the GloSea5 forecasts using the RAFT algorithm. This results in a climatological forecast with a structure comparable to an autoregressive process of order one. The updated climatological forecast with lead time of one week is indicated with a brown line in Figure 4. This forecast has 5-10% higher skill than the climatological reference forecast. Furthermore, a forecast that combines the lagged ensemble means post-processed with RAFT is the best forecast for weeks 6 and 7, and comparable to the RAFT climatology for week 8 and onward.

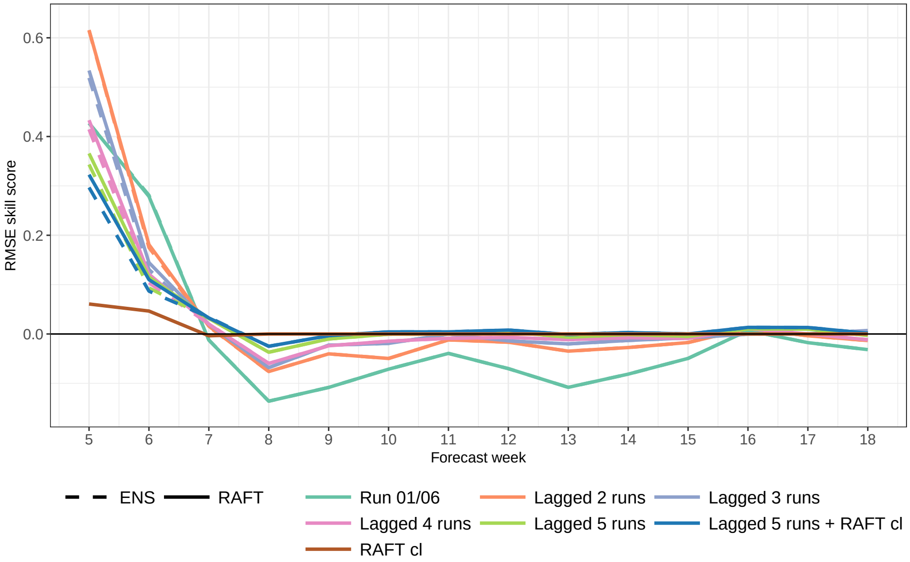


Figure 5. Root mean squared error (RMSE) skill scores for various model combinations for forecasts issued in week five compared against ERA5 climatology. Each run combination consists of the most recently available runs. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015.

For a further comparison of various model combinations in an operational setting, Figure 5 shows the skill scores for a number of forecasts for weeks 5-18 issued in week 5. These results indicate that an optimal forecasting strategy is to combine a smaller number of the most recent runs for the first two forecast weeks after which all five runs as well as the climatology should be combined. While the combination of all runs and climatology does not outperform climatology for all forecasts weeks, it is overall the best forecast for weeks 7-18. In particular, including climatology in the ensemble is consistently slightly better than only considering the five GloSea5 runs.

As shown in Figure 2, the adjustment period at forecast lead time 5 is relatively short. This can also be seen in Figure 5 where the RAFT-adjusted forecasts coincide with the original forecasts from week 7. Figure 6 shows the same original forecasts from week 10 and onward, as well as the RAFT-adjusted forecasts issued in week 10. Here, the adjustment periods are considerably longer for all the forecast runs and the RAFT adjustment yields improved performance until week 14 after which the forecasts again coincide. In this case, the original forecasts have lead times of 5+ weeks, and we see that only the combination of all five runs is on a par with climatology. For a combination of three or more runs, the RAFT adjustment yields an overall higher skill than climatology with the full combination of all five runs and RAFT climatology again showing the highest skill overall.

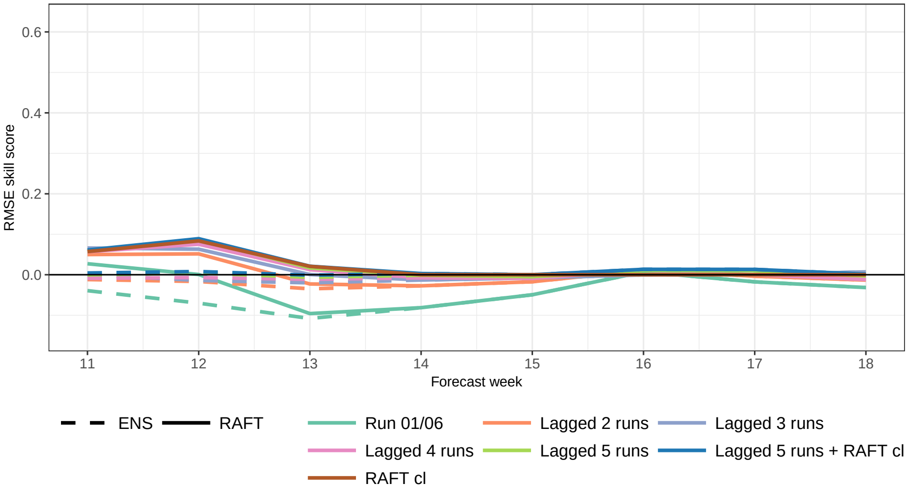


Figure 6. Root mean squared error (RMSE) skill scores for various model combinations compared against ERA5 climatology for forecast weeks 10-18. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015. The original GloSea5 forecasts are indicated with dashed lines while RAFT forecasts issued in week 10 are indicated by solid lines.

4 Conclusions and discussion

In a study of long-range forecast skill for weekly summer surface temperatures in Europe, we assess the skill of the UK Met Office's seasonal prediction system GloSea5 against the ERA5 reanalysis. GloSea5 uses a lagged initialization approach where, for the 1993-2015 hindcasts analyzed here, seven members are initialized on the 1st, 9th, 17th and 25th of every month. Our results indicate that the system might benefit from a step-wise model combination approach, where for the earliest forecast lead times, only more recently available runs are used, while a larger set of runs should be employed for lead times beyond two weeks. Furthermore, the forecast skill is increased for lead time beyond two weeks if climatology is included in the ensemble.

For a lagged ensemble system, additional information in the form of observed forecast errors is available for earlier lead times of the older ensemble members. Using the recently proposed RAFT adjustment approach (Schuhen et al., 2020), we have investigated the use of this information to post-process the older members before the forecast is issued. Our results indicate that the application of the RAFT adjustment can improve the RMSE skill of the forecast by as much as 10% compared to climatology. In each time step, the length of the RAFT adjustment period depends on the number of future lead times where the forecast error is expected to correlate with the most recently observed forecast error. We find that the length of the adjustment period varies over time, with a higher correlation across lead times in July and August than in the earlier part of our study period in May and June.

As argued by e.g. Kharin and Zwiers (2003) and Van Schaeybroeck and Vannitsem (2018), the small samples sizes available for seasonal forecasts (23 seasons in our case) require simple post-processing methods in order to avoid overfitting. The RAFT approach is a fairly simple post-processing method whose strength lies in the use of new, otherwise unused, information. The current study focuses on average skill in predicting mean weekly summer temperatures in Europe. For many forecast users, a particularly valuable information is the occurrence of outliers, e.g. a particularly warm or cold summer. While this topic requires further investigation, we expect that RAFT could prove particularly useful in such situations when the outlier has been detected in the newest runs with that not being the case for the older runs.

Acknowledgments

The authors acknowledge the support of the Research Council of Norway: N. Schuhen through grant nr. 259864 “Stipendiatstillinger til Norsk Regnesentral”, T. L. Thorarinsdottir and A. Lenkoski through grant nr. 270733 “Seasonal Forecasting Engine”.

References

Copernicus Climate Change Service (C3S) (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store. https://cds.climate.copernicus.eu/cdsapp#!/home, accessed in November 2019. 5

Doblas-Reyes, F., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L., and Van Oldenborgh, G. (2013a). Initialized near-term regional climate change prediction. *Nature Communications*, 4:1715. 4

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R.

(2013b). Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, 4(4):245–268. 4

Heinrich, C., Hellton, K. H., Lenkoski, A., and Thorarinsdottir, T. L. (2019). Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. arXiv:1907.09716. 4

Hoskins, B. (2013). The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, 139(672):573–584. 4

Kharin, V. V. and Zwiers, F. W. (2003). Improved seasonal probability forecasts. *Journal of Climate*, 16(11):1684–1701. 11

MacLachlan, C., Arribas, A., Peterson, K., Maidens, A., Fereday, D., Scaife, A., Gordon, M., Vellinga, M., Williams, A., Comer, R., Camp, J., Xavier, P., and Madec, G. (2015). Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, 141(689):1072–1084. 4, 5

Ogallo, L., Bessemoulin, P., Ceron, J.-P., Mason, S., and Connor, S. J. (2008). Adapting to climate variability and change: the climate outlook forum process. *Bulletin of the World Meteorological Organization*, 57(2):93–102. 4

Robertson, A. and Vitart, F., editors (2018). *Sub-seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*. Elsevier, Amsterdam, Netherlands. 4

Roulin, E. and Vannitsem, S. (2019). Post-processing of seasonal predictions – case studies using EUROSIP hindcast data base. *Nonlinear Processes in Geophysics*, in review. 4

Royer, J. (1993). Review of recent advances in dynamical extended range forecasting for the extratropics. In Shukla, J., editor, *Prediction of Interannual Climate Variations*, pages 49–69. Springer, Berlin, Heidelberg. 4

Schuhen, N. (2019). Order of operation for multi-stage post-processing of ensemble wind forecast trajectories. *Nonlinear Processes in Geophysics*, accepted. 7

Schuhen, N., Thorarinsdottir, T. L., and Lenkoski, A. (2020). Rapid adjustment and post-processing of temperature forecast trajectories. *Quarterly Journal of the Royal Meteorological Society*, in press. 4, 6, 7, 11

Thorarinsdottir, T. L. and Schuhen, N. (2018). Verification: assessment of calibration and accuracy. In Vannitsem, S., Wilks, D. S., and Messner, J. W., editors, *Statistical postprocessing of ensemble forecasts*, chapter 6, pages 155–186. Elsevier, Amsterdam, Netherlands. 5

Van Schaeybroeck, B. and Vannitsem, S. (2018). Postprocessing of long-range forecasts. In Vannitsem, S., Wilks, D. S., and Messner, J. W., editors, *Statistical Postprocessing of Ensemble Forecasts*, chapter 10, pages 267–290. Elsevier, Amsterdam, Netherlands. 4, 11

Vitart, F., Robertson, A. W., and Anderson, D. L. (2012). Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *Bulletin of the World Meteorological Organization*, 61(2):23. 4

Paper IV

Verification: assessment of calibration and accuracy

Thordis L. Thorarinsdottir, Nina Schuhen

In: *Statistical postprocessing of ensemble forecasts*.
Ed. by Vannitsem, S., Wilks D. S. and Messner, J. W. Elsevier.
Chap. 6, pp. 155–186, 2018.
DOI: 10.1016/b978-0-12-812372-0.00006-6.

Statistical Postprocessing of Ensemble Forecasts

Edited by

Stéphane Vannitsem
Royal Meteorological Institute of Belgium, Brussels, Belgium

Daniel S. Wilks
Cornell University, Ithaca, NY, United States

Jakob W. Messner
Technical University of Denmark, Kongens Lyngby, Denmark

ELSEVIER

CHAPTER

VERIFICATION: ASSESSMENT OF CALIBRATION AND ACCURACY 6

Thordis L. Thorarinsdottir, Nina Schuhen
Norwegian Computing Center, Oslo, Norway

CHAPTER OUTLINE

6.1 Introduction 155
6.2 Calibration 156
6.2.1 Univariate Calibration 157
6.2.2 Multivariate Calibration 158
6.2.3 Example: Comparing Multivariate Ranking Methods 159
6.3 Accuracy 161
6.3.1 Univariate Assessment 161
6.3.2 Simulation Study: Comparing Univariate Scoring Rules 165
6.3.3 Assessing Extreme Events 170
6.3.4 Example: Proper and Nonproper Verification of Extremes 174
6.3.5 Multivariate Assessment 175
6.3.6 Divergence Functions 177
6.3.7 Testing Equal Predictive Performance 179
6.4 Understanding Model Performance 180
6.5 Summary 182
References 183

6.1 INTRODUCTION

In a discussion article on the application of mathematics in meteorology, Bigelow (1905) describes the fundamentals of modeling in a timeless manner:

> There are three processes that are generally essential for the complete development of any branch of science, and they must be accurately applied before the subject can be considered to be satisfactorily explained. The first is the discovery of a mathematical analysis, the second is the discussion of numerous observations, and the third is a correct application of the mathematics to the observations, including a demonstration that these are in agreement.

Statistical Postprocessing of Ensemble Forecasts. https://doi.org/10.1016/B978-0-12-812372-0.00006-6

The topic of this chapter is methods for carrying out the last item on Bigelow's list, that is, methods to demonstrate the agreement between a model and a set of observations. Ensemble prediction systems and statistically postprocessed ensemble forecasts provide probabilistic predictions of future weather. Verification methods applied to these systems should thus be equipped to handle both the verification of the best prediction derived from the ensemble and the verification of the associated prediction uncertainty.

Murphy (1993) argues that a general prediction system should strive to perform well on three types of goodness: There should be consistency between the forecaster's judgment and the forecast, there should be correspondence between the forecast and the observation, and the forecast should be informative for the user. Similarly, Gneiting, Balabdaoui, and Raftery (2007) state that the goal of probabilistic forecasting should be to maximize the sharpness of the predictive distribution subject to calibration. Here, calibration refers to the statistical consistency between the forecast and the observation, while sharpness refers to the concentration of the forecast uncertainty; the sharper the forecast, the higher information value will it provide, as long as it is also calibrated. The prediction goal of Gneiting et al. (2007) is thus equivalent to Murphy's second and third types of goodness.

We focus on verification methods for probabilistic predictions of continuous variables in one or more dimensions under the general framework described by Murphy (1993) and Gneiting et al. (2007). Specifically, we denote an observation in d dimensions by $y = (y_1, \ldots, y_d) \in \Omega^d$ for $d = 1, 2, \ldots$, where Ω denotes either the real axis $\mathbb{R}$, the nonnegative real axis $\mathbb{R}_{\geq 0}$, the positive real axis $\mathbb{R}_{>0}$, or an interval on $\mathbb{R}$. A probabilistic forecast for y given by a distribution function with support on Ω^d is denoted by $F \in \mathcal{F}$ for some appropriate class of distributions $\mathcal{F}$, with the density denoted by f if it exists. For ensemble forecasts, we will alternatively use the notation $\mathbf{x} = \{x_1, \ldots, x_K\}$ to describe the K ensemble members or F for the associated empirical distribution function. Verification methods for deterministic predictions and other types of variables are discussed, for example, in Wilks (2011, Chapter 8) and Jolliffe and Stephenson (2012).

This chapter is organized as follows. Diagnostic tools for checking calibration are discussed in Section 6.2. Section 6.3 describes methods that assess the accuracy of forecasts where each forecast is issued a numerical score based on the event that materializes. Scoring rules apply to individual events while divergence functions compare the empirical distribution of a series of events with a predictive distribution. The scores may focus on certain aspects of the forecast, such as the tails, and it is important also to assess the uncertainty in the scores. The properties of various univariate scores are compared in a simulation study. While the methods in Section 6.3 provide a decision-theoretically coherent approach to model evaluation and model ranking, they may hide key information about the model performance such as the direction of bias. Additional evaluation may thus be needed to better understand the performance of a single model. Approaches for this are discussed in Section 6.4. The chapter then closes with a summary in Section 6.5.

6.2 CALIBRATION

Calibration, or reliability, is the most fundamental aspect of forecast skill for probabilistic forecasts as it is a necessary condition for the optimal use and value of the forecast. Calibration refers to the statistical compatibility between the forecast and the observation; the forecast is calibrated if the observation cannot be distinguished from a random draw from the predictive distribution.

6.2.1 UNIVARIATE CALIBRATION

Several alternative notions of univariate calibration exist for a single forecast (Gneiting et al., 2007; Tsyplakov, 2013) and a group of forecasts (Strähl & Ziegel, 2017). We focus on the so-called *probabilistic calibration* as suggested by Dawid (1984); F is probabilistically calibrated if the *probability integral transform* (PIT) $F(Y)$, the value of the predictive cumulative distribution function for the random observation Y, is uniformly distributed. If F has a discrete component, a randomized version of the PIT given by

$$\lim_{y \uparrow Y} F(y) + V\left(F(Y) - \lim_{y \uparrow Y} F(y)\right)$$

with $V \sim \mathcal{U}([0,1])$ may be used, see Gneiting and Ranjan (2013). Here, we use $y \uparrow Y$ to denote that the limit is taken as y approaches Y from below.

Assume our test set consists of n observations $y_1, \ldots, y_n$. For a forecasting method issuing continuous univariate predictive distributions $F_1, \ldots, F_n$, calibration can be assessed empirically by plotting the histogram of the PIT values

$$F_1(y_1), \ldots, F_n(y_n).$$

A forecasting method that is calibrated on average will return a uniform histogram, a ∩-shape indicates overdispersion and a ∪-shape indicates underdispersion, while a systematic bias results in a triangular-shaped histogram. Examples of miscalibration are shown in Fig. 6.1, including a biased forecast (panel a), an underdispersive forecast (panel b), an overdispersive forecast (panel c), and an example of a multiply misspecified forecast where the left tail is too light, the main bulk of the distribution lacks mass and the right tail is too heavy (panel d).

The discrete equivalent of the PIT histogram, which applies to ensemble forecasts, is the verification rank histogram (Anderson, 1996; Hamill & Colucci, 1997). It shows the distribution of the ranks of the observations within the corresponding ensembles and has the same interpretation as the PIT histogram.

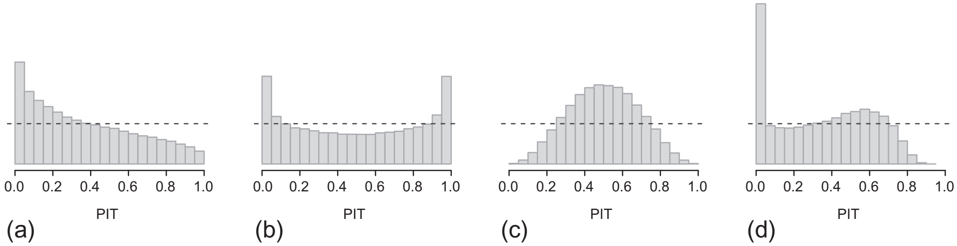


FIG. 6.1

Probability integral transform (PIT) histograms for 100,000 simulated standard Gaussian $\mathcal{N}(0,1)$ observations and various misspecified forecasts: (a) biased $\mathcal{N}(0.5,1)$ forecasts, (b) underdispersive $\mathcal{N}(0,0.75^2)$ forecasts, (c) overdispersive $\mathcal{N}(0,2^2)$ forecasts, and (d) multiply misspecified generalized extreme value GEV(0, 1, 0.5) forecasts. The theoretically optimal histograms are indicated with *dashed lines*.

The information provided by a rank histogram may also be summarized numerically by the reliability index (RI), which is defined as

$$\mathrm{RI} = \sum_{i=1}^{I} \left| \zeta_i - \frac{1}{I} \right|$$

where I is the number of (equally sized) bins in the histogram and ζ_i is the observed relative frequency in bin $i = 1, \ldots, I$. The RI thus measures the departure of the rank histogram from uniformity (Delle Monache, Hacker, Zhou, Deng, & Stull, 2006).

6.2.2 MULTIVARIATE CALIBRATION

For assessing the calibration of multivariate forecasts, Gneiting, Stanberry, Grimit, Held, and Johnson (2008) formalized a general two-step framework. Let $S = \{x_1, \ldots, x_K, y\}$ denote a set of $K+1$ points in Ω^d comprising an ensemble forecast with K members and the corresponding observation y. The rank of y in S, $\mathrm{rank}_S(y)$, is calculated in two steps,

(i) apply a prerank function $\rho_S : \Omega^d \to \mathbb{R}_{\geq 0}$ to calculate the prerank $\rho_S(u)$ of every $u \in S$ resulting in a univariate value for each u;

(ii) set the rank of the observation y equal to the rank of $\rho_S(y)$ in $\{\rho_S(x_1), \ldots, \rho_S(x_K), \rho_S(y)\}$,

$$\mathrm{rank}_S(y) = \sum_{v \in S} \mathbb{1}\{\rho_S(v) \leq \rho_S(y)\}$$

where $\mathbb{1}$ denotes the indicator function and ties are resolved at random.

Here, we focus on four different approaches that follow this general two-step framework. Further approaches are discussed in Gneiting et al. (2008), Ziegel and Gneiting (2014), and Wilks (2017). The difference between our four approaches lies in the definition of the prerank function ρ_S in step (i). The *multivariate ranking* of Gneiting et al. (2008) is defined using the prerank function

$$\rho_S^{\mathrm{m}}(u) = \sum_{v \in S} \mathbb{1}\{v \preccurlyeq u\} \tag{6.1}$$

where $v \preccurlyeq u$ if and only if $v_i \leq u_i$ in all components $i = 1, \ldots, d$. Gneiting et al. (2008) further consider an optional initial step in the ranking procedure in which the data is normalized in each component before the ranking. The *average ranking* proposed by Thorarinsdottir, Scheuerer, and Heinz (2016) provides a similar ascending rank structure and is given by the average over the univariate ranks. That is, let

$$\mathrm{rank}_S(u, i) = \sum_{v \in S} \mathbb{1}\{v_i \leq u_i\}$$

denote the standard univariate rank of the ith component of u among the values in S. The multivariate average rank is then defined using the prerank function

$$\rho_S^{\mathrm{a}}(u) = \frac{1}{d} \sum_{i=1}^{d} \mathrm{rank}_S(u, i) \tag{6.2}$$

Two further approaches assess the centrality of the observation within the ensemble. Under *minimum spanning tree ranking*, the prerank function $\rho_S^{\mathrm{mst}}(u)$ is given by the length of the minimum spanning tree of the set $S \setminus u$, that is, the set S without the element u (Smith & Hansen, 2004; Wilks, 2004). Here, a spanning tree of the set $S\backslash u$ is a collection of $K - 1$ edges such that all points in $S\backslash u$ are used, with no closed loops. The spanning tree with the smallest length is then the minimum spanning tree (Kruskal, 1956); it may, for example, be calculated using the `R` package `vegan` (Oksanen et al., 2017; R Core Team, 2016).

Alternatively, the *band-depth ranking* proposed by Thorarinsdottir et al. (2016) uses a prerank function that calculates the proportion of components of $u \in S$ inside bands defined by pairs of points from S. It can be written as

$$\rho_S^{\mathrm{bd}}(u) = \frac{1}{d}\sum_{i=1}^{d}\left[\mathrm{rank}_S(u,i)[(K+1) - \mathrm{rank}_S(u,i)] + [\mathrm{rank}_S(u,i) - 1]\sum_{v\in S}\mathbb{1}\{v_i = u_i\}\right] \tag{6.3}$$

If $u_i \neq v_i$ with probability 1 for all $u, v \in S$ with $u \neq v$ and $i = 1, \ldots, d$ the formula in Eq. (6.3) may be simplified to

$$\rho_S^{\mathrm{bd}}(u) = \frac{1}{d}\sum_{i=1}^{d}[(K+1) - \mathrm{rank}_S(u,i)][\mathrm{rank}_S(u,i) - 1] \tag{6.4}$$

This implies that the formula in Eq. (6.3) should be used for forecasts with a discrete component, for example, precipitation forecasts. The band depth in Eq. (6.3) is equivalent to the simplicial depth proposed by Liu (1990) and thus also to the simplicial depth ranking proposed by Mirzargar and Anderson (2017), see López-Pintado and Romo (2009) and Thorarinsdottir et al. (2016).

While all four methods return a uniform rank histogram for a calibrated forecast, the interpretation of the histogram shape for a misspecified forecast varies between the methods as demonstrated in the following example.

6.2.3 EXAMPLE: COMPARING MULTIVARIATE RANKING METHODS

The four multivariate ranking methods are compared in Fig. 6.2 for several different settings where $y \in \mathbb{R}^d$ can be thought of as a temporal trajectory of a real-valued variable observed at $d = 10$ equidistant time points $t = 1, \ldots, 10$. In the first two examples (rows 1 and 2), y is a realization of a zero-mean Gaussian AR(1) (autoregressive) process Y with a covariance function given by

$$\mathrm{Cov}(Y_i, Y_j) = \exp(-|i-j|/\tau), \quad \tau > 0. \tag{6.5}$$

The process Y thus has standard Gaussian marginal distributions while the parameter τ controls how fast correlations decay with time lag. We set $\tau = 3$ for Y and consider ensemble forecasts with 50 members of the same type, but with a different parameter value τ. That is, we set $\tau = 1.5$ in row 1 (too strong correlation) and $\tau = 5$ in row 2 (too weak correlation). It follows from this construction that a univariate calibration test at a fixed time point would not detect any miscalibration in the forecasts.

While all four methods are able to detect the misspecification in the correlation structure, the resulting histograms vary in shape. The shape of the average rank histograms and the band-depth rank histograms offer a similar interpretation as that of the univariate rank histograms in Fig. 6.1 with a ∪-shape when the correlation is too strong (underdispersion across components) and a ∩-shape when the correlation is too weak (overdispersion across components). In these 10-dimensional examples, the prerank ordering of the multivariate rank histograms (Eq. 6.1) is only able to detect miscalibration related

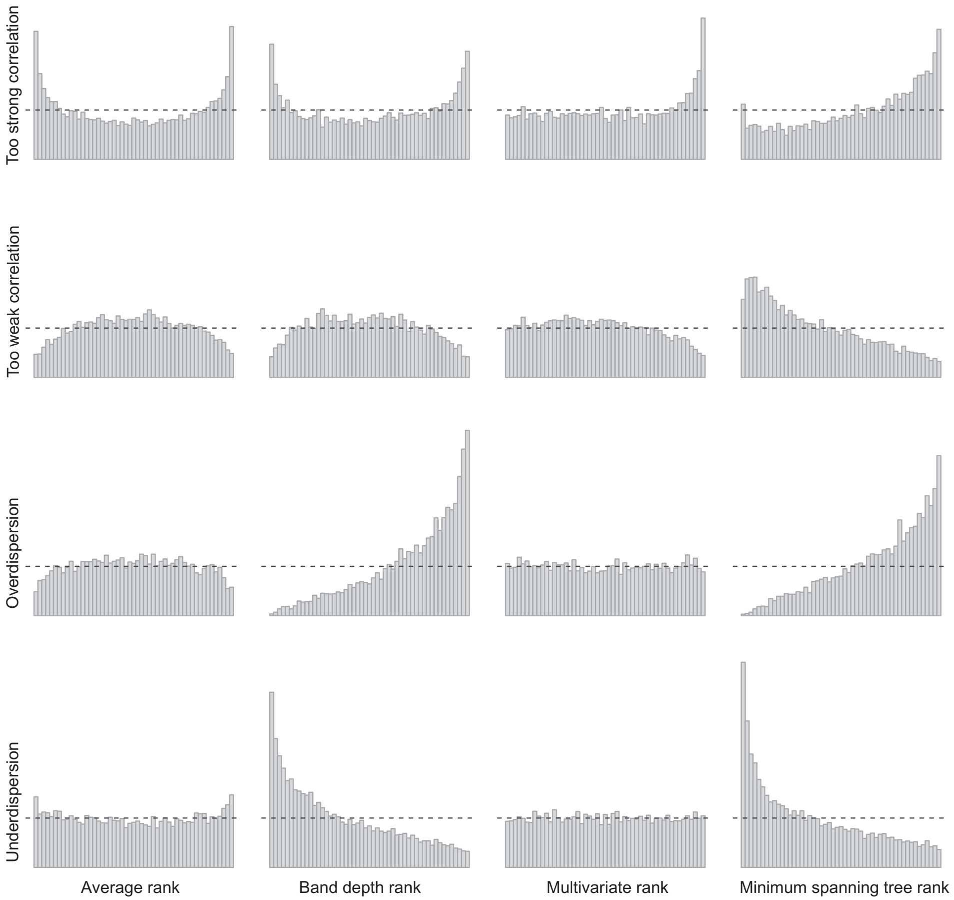


FIG. 6.2

Rank histograms for multivariate data showing various types of miscalibration under different ranking methods: average ranking (*first column*), band-depth ranking (*second column*), multivariate ranking (*third column*), and minimum spanning tree ranking (*fourth column*); 10,000 simulated observations of dimension 10 are compared with ensemble forecasts with 50 members. In the *top two rows*, the observations are realizations of a zero-mean Gaussian AR(1) process with the covariance function in Eq. (6.5) where $\tau = 3$. The forecasts follow the same model with $\tau = 1.5$ (*first row*) and $\tau = 5$ (*second row*). In the *bottom two rows*, the observations are i.i.d. standard Gaussian variables while the forecasts have variance 1.25^2 (*third row*) and 0.85^2 (*fourth row*). The theoretically optimal histograms are indicated with *dashed lines*.

to the highest ranks (see also the discussion in Pinson & Girard, 2012 and Thorarinsdottir et al., 2016). Under minimum spanning tree ranking, too many observations have high ranks when the correlation in the forecasts is too strong and the opposite holds for the example with too weak correlation in the forecasts.

In the latter two examples in Fig. 6.2 (rows 3 and 4), both observations and forecasts are i.i.d. variables in 10 dimensions. However, the marginal distributions of the ensemble forecasts are misspecified. The observations follow a standard Gaussian distribution, the forecasts in row 3 have a standard deviation of 1.25 (overdispersion) and the forecasts in row 4 have a standard deviation of 0.85 (underdispersion). The shape of the average rank histograms is exactly that of their univariate counterparts in Fig. 6.1, indicating that this ranking method cannot distinguish between miscalibration in the marginals and the higher-order structure. For the two ranking methods based on centrality, the marginal overdispersion results in too many high ranks while the marginal underdispersion results in too many low ranks. For this dimensionality, the multivariate ranking is unable to detect the miscalibration.

Further comparison of the four ranking methods is provided in Thorarinsdottir et al. (2016) and Wilks (2017). In general, it is a challenging task to represent and compare a multifaceted higher-order structure with a single value. As the different methods vary in their strengths and weaknesses, it is recommended that several of these methods be applied when assessing multivariate calibration. The multivariate ranking of Gneiting et al. (2008), for instance, does not satisfy affine invariance (Mirzargar & Anderson, 2017) while lower-dimensional positive and negative biases may cancel out under average ranking (Thorarinsdottir et al., 2016).

Furthermore, a prior assessment of the marginal calibration may increase the information value in the multivariate rank histograms and ease the interpretation of the resulting shapes. As the multivariate methods perform a simultaneous assessment of the marginal and the higher-order calibration, a specific nonuniform shape may represent multiple types of misspecifications. For example, depth-based approaches such as the band-depth ranking and the minimum spanning tree ranking are not able to distinguish between underdispersive and biased forecasts (Mirzargar & Anderson, 2017).

6.3 ACCURACY

In this section, we discuss methods for assessing forecast accuracy that are appropriate for ranking and comparing competing forecasting methods. Alternative assessment techniques that may provide additional insights for understanding the performance and errors of a single forecasting model, but are not appropriate for forecast ranking are discussed in Section 6.4.

6.3.1 UNIVARIATE ASSESSMENT

Scoring rules assess the accuracy of probabilistic forecasts by assigning a numerical penalty to each forecast-observation pair. Specifically, a scoring rule is a mapping

$$S : \mathcal{F} \times \Omega^d \to \mathbb{R} \cup \{\infty\} \tag{6.6}$$

where for every $F \in \mathcal{F}$ the map $y \mapsto S(F, y)$ is quasiintegrable. In our notation, a smaller penalty indicates a better prediction. A scoring rule is *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_G S(G,Y) \leq \mathbb{E}_G S(F,Y) \tag{6.7}$$

for all probability distributions $F, G \in \mathcal{F}$, that is, if the expected score for a random observation Y is optimized if the true distribution of Y (G) is issued as the forecast. The scoring rule is *strictly proper* relative to the class $\mathcal{F}$ if Eq. (6.7) holds with equality only if $F = G$. Propriety will encourage honesty and prevent hedging, which coincides with Murphy's first type of goodness (Murphy, 1993). That is, the scores cannot be hedged by a willful divergence of the forecast from the true distribution to improve the perceived performance, see for example the discussion in Section 1 of Gneiting (2011).

Competing forecasting methods are verified based on a proper scoring rule by comparing their mean scores over an out-of-sample test set. The method with the smallest mean score is preferred. Formal tests of the null hypothesis of equal predictive performance can also be employed, see Section 6.3.7. While average scores are directly comparable if they refer to the same set of forecast situations, this may no longer hold for distinct sets of forecast cases, for instance due to spatial and temporal variability in the predictability of weather. For ease of interpretability and to address this issue, verification results are sometimes represented as a *skill score* of the form

$$S_n^{\text{skill}}(A) = \frac{\frac{1}{n}\sum_{i=1}^{n} S(F_i^A, y_i) - \frac{1}{n}\sum_{i=1}^{n} S(F_i^{\text{ref}}, y_i)}{\frac{1}{n}\sum_{i=1}^{n} S(F_i^{\text{perf}}, y_i) - \frac{1}{n}\sum_{i=1}^{n} S(F_i^{\text{ref}}, y_i)} \tag{6.8}$$

for the forecasting method A where F^{ref} denotes the forecast from a reference method, F^{perf} denotes the perfect forecast, and n is the size of the test set. The skill score is standardized such that it takes the value 1 for an optimal forecast and the value 0 for the reference forecast. Negative values thus indicate that the forecasting method A is of a lesser quality than the reference forecast. However, it is vital to select the reference forecast with care (Murphy, 1974, 1992) as skill scores of the form of Eq. (6.8) may be improper even if the underlying scoring rule S is proper (Gneiting & Raftery, 2007; Murphy, 1973a).

The most popular proper scoring rules for univariate real-valued quantities are the *ignorance* (or *logarithmic*) *score* (IGN) and the continuous ranked probability score, see Gneiting and Raftery (2007) for a more comprehensive list. IGN is defined as

$$\text{IGN}(F, y) = -\log f(y) \tag{6.9}$$

where f denotes the density of F (Good, 1952). It thus applies to absolutely continuous distributions only and cannot be applied directly to ensemble forecasts. For a large enough ensemble, the density of the ensemble forecast may potentially be approximated using, for example, kernel density estimation or by fitting a parametric distribution. Alternatively, IGN may be replaced by the *Dawid-Sebastiani* (DS) *score* (Dawid & Sebastiani, 1999),

$$\text{DS}(F, y) = \log \sigma_F^2 + \frac{(y - \mu_F)^2}{\sigma_F^2} \tag{6.10}$$

where μ_F denotes the mean value of F and σ_F^2 its variance. While the proper DS score equals IGN for a Gaussian predictive distribution F, it only requires the estimation of the ensemble mean and variance.

The *continuous ranked probability score* (CRPS) (Matheson & Winkler, 1976) is of particular interest in that it simultaneously assesses both calibration and sharpness, and thus all three types of goodness discussed by Murphy (1993). The CRPS applies to probability distributions with a finite mean and has three equivalent definitions (Gneiting & Raftery, 2007; Gneiting & Ranjan, 2011; Hersbach, 2000; Laio & Tamea, 2007),

$$\mathrm{CRPS}(F,y) = \mathbb{E}_F|X-y| - \frac{1}{2}\mathbb{E}_F\mathbb{E}_F|X-X'| \tag{6.11}$$

$$= \int_{-\infty}^{+\infty} (F(x) - \mathbb{1}\{y \le x\})^2 \mathrm{d}x \tag{6.12}$$

$$= \int_0^1 \left(F^{-1}(\tau) - y\right)\left(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau\right)\mathrm{d}\tau \tag{6.13}$$

Here, X and X' denote two independent random variables with distribution F, $\mathbb{1}\{y \le x\}$ denotes the indicator function that is equal to 1 if $y \le x$ and 0 otherwise, and $F^{-1}(\tau) = \inf\{x \in \mathbb{R} : \tau \le F(x)\}$ is the quantile function of F.

It follows directly from Eqs. (6.12), (6.13) that the CRPS is tightly linked to other proper scores that focus on specific parts of the predictive distribution. The form in Eq. (6.12) can be interpreted as the integral over the *Brier score* (Brier, 1950), which assesses the predictive probability of threshold exceedance. The Brier score is usually written in the form

$$\mathrm{BS}(F,y|u) = (p_u - \mathbb{1}\{y \ge u\})^2 \tag{6.14}$$

for a threshold u with $p_u = 1 - F(u)$. Similarly, the integrand in Eq. (6.13) equals the *quantile score* (Friederichs & Hense, 2007; Gneiting & Raftery, 2007),

$$\mathrm{QS}(F,y|q) = (F^{-1}(q) - y)(\mathbb{1}\{y \le F^{-1}(q)\} - q) \tag{6.15}$$

which assesses the predicted quantile $F^{-1}(q)$ for a probability level $q \in (0, 1)$.

When the predictive distribution F is given by a finite ensemble $\{x_1, \ldots, x_K\}$, the CRPS representation in Eq. (6.11) is equal to

$$\mathrm{CRPS}(F,y) = \frac{1}{K}\sum_{k=1}^{K}|x_k - y| - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}|x_k - x_l| \tag{6.16}$$

see Grimit, Gneiting, Berrocal, and Johnson (2006). For small ensembles, Ferro, Richardson, and Weigel (2008) propose a *fair* approximation given by

$$\mathrm{CRPS}(F,y) \approx \frac{1}{K}\sum_{k=1}^{K}|x_k - y| - \frac{1}{2K(K-1)}\sum_{k=1}^{K}\sum_{l=1}^{K}|x_k - x_l| \tag{6.17}$$

For large ensembles, a more computationally efficient calculation is based on the generalized quantile function (Laio & Tamea, 2007). Let $x_{(1)} \le \cdots \le x_{(K)}$ denote the order statistics of $x_1, \ldots, x_K$. Then

$$\mathrm{CRPS}(F,y) = \frac{2}{K^2}\sum_{i=1}^{K}(x_{(i)} - y)\left(K\mathbb{1}\{y < x_{(i)}\} - i + \frac{1}{2}\right) \tag{6.18}$$

see also Murphy (1970). The formula in Eq. (6.18) is implemented in the `R` package `scoringRules` together with exact formulas for a large class of parametric families of distributions (see Table 6.1 and Jordan, Krüger, & Lerch, 2017).

Table 6.1 Parametric Families of Distributions for Which the CRPS Is Implemented in the R Package scoringRules (Jordan et al., 2017)

Dist. on $\mathbb{R}$	Dist. on $\mathbb{R}_{>0}$	Dist. on Intervals	Discrete Dist.
Gaussian	Exponential	Generalized extreme value	Poisson
t	Gamma	Generalized Pareto	Neg. binomial
Logistic	Log-Gaussian	Trunc. Gaussian	
Laplace	Log-logistic	Trunc. *t*	
Two-piece Gaussian	Log-Laplace	Trunc. logistic	
Two-piece exponential		Trunc. exponential	
Mixture of Gaussians		Uniform	
		Beta	

Notes*: The truncated families can be defined with or without a point mass at the support boundaries.*

When the forecasting model is estimated using a Bayesian analysis, the predictive distribution F is commonly given by the posterior predictive distribution under the model. Here, F is rarely known in closed form and is, instead, approximated by a large sample that is often obtained using Markov chain Monte Carlo techniques. However, such techniques may yield highly correlated samples, which complicates the employment of approximation formulas as those for the CRPS shown herein. Optimal approximations for both IGN and CRPS when the distribution F is the posterior predictive distribution from a Bayesian analysis are discussed in Krüger, Lerch, Thorarinsdottir, and Gneiting (2016).

The quality of a deterministic forecast x is typically assessed by applying a *scoring function* $s(x, y)$, that assigns a numerical score based on x and the corresponding observation y. As in the case of proper scoring rules, competing forecasting methods are compared and ranked in terms of the mean scores over the cases in a test set. Popular scoring functions include the squared error, $s(x, y) = (x - y)^2$, and the absolute error, $s(x, y) = |x - y|$.

A scoring function can be applied to a probabilistic prediction $F \in \mathcal{F}$ if it is *consistent* for a functional T relative to the class $\mathcal{F}$ in the sense that

$$\mathbb{E}_F s(T(F), Y) \leq \mathbb{E}_F s(x, Y) \tag{6.19}$$

for all $x \in \Omega$ and $F \in \mathcal{F}$. A consistent scoring function becomes a proper scoring rule if the functional T in Eq. (6.19) is used as the derived deterministic prediction based on F. That is, if $S(F, y) = s(T(F), y)$. The squared error proper scoring rule is given by

$$\text{SE}(F, y) = (\text{mean}(F) - y)^2 \tag{6.20}$$

where mean(F) denotes the mean value of F, and the absolute error proper scoring rule becomes

$$\text{AE}(F, y) = |\text{med}(F) - y| \tag{6.21}$$

where med(F) denotes the median of F.

One appealing property of scoring rules that derive from scoring functions is thus the possibility of comparing deterministic and probabilistic forecasts. See Gneiting (2011) for an extensive discussion of the use of scoring functions to evaluate probabilistic predictions.

6.3.2 SIMULATION STUDY: COMPARING UNIVARIATE SCORING RULES

The purpose of this simulation study is to demonstrate a coherent approach to using proper scores and rank or PIT histograms in practice, while highlighting some of the difficulties that might arise when working with limited data sets. In particular, we investigate how different scoring rules rank forecasts according to their skill, and how these results differ with the amount of available data.

We start by generating two sets of observation data, drawn randomly from the same fixed “true” distribution. The first set consists of 100 values, which will serve as verifying observations, while the second set, the training data, consists of 300 values for each of the 100 observations. Our goal is to issue forecasts matching the observations, based on the information contained in the training data. For the first part of the simulation study, the true distribution is normal, with a random mean $\mu \sim \mathcal{N}(25,1)$ and fixed standard deviation $\sigma = 3$. In the second part, the truth is a Gumbel distribution, with the mean following a $\mathcal{N}(25,1)$ distribution and the scale parameter fixed to 3, see Table 6.2.

Using a method-of-moments approach, we estimate four competing forecast distributions for each observation, which are listed in Table 6.3. The distribution parameters are calculated by plugging the sample mean and sample standard deviation from the training data into the equations for mean and variance. For the noncentral t-distribution, the degrees of freedom are obtained numerically by a root-finding algorithm described in Brent (1973), while restricting them to $\nu \geq 3$, ensuring that both mean and variance exist. As a fifth forecaster, we use the true distribution, from which the observations

Table 6.2 Observation-Generating Distributions Used in the Simulation Study

	Distribution $F(Y)$		$\mathbb{E}(Y)$	$\mathrm{Var}(Y)$
Part 1	Normal	$\mathcal{N}(\mu,\sigma^2)$	$\mu \sim \mathcal{N}(25,1)$	$\sigma^2 = 9$
Part 2	Gumbel	$\mathrm{G}(\mu,\sigma)$	$\mu + \sigma \cdot \gamma \sim \mathcal{N}(25,1)$	$\frac{\pi^2}{6}\sigma^2 = \frac{3\pi^2}{2}$

Notes*: The expected values are random variables following a normal distribution, while the scale parameters are fixed. γ denotes the Euler-Mascheroni constant.*

Table 6.3 Forecasters Used in Both Parts of the Simulation Study, and Their Expected Values and Variances as Functions of the Distribution Parameters

Distribution $F(Y)$		$\mathbb{E}(Y)$	$\mathrm{Var}(Y)$
Normal	$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
Noncentral t	$t(\nu,\mu)$	$\mu\sqrt{\frac{\nu}{2}}\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$, if $\nu > 1$	$\frac{\nu(1+\mu^2)}{\nu-2} - \frac{\mu^2\nu}{2}\left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2$, if $\nu > 2$
Lognormal	$\ln\mathcal{N}(\mu,\sigma^2)$	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$	$(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$
Gumbel	$\mathrm{G}(\mu,\sigma)$	$\mu + \sigma \cdot \gamma$	$\frac{\pi^2}{6}\sigma^2$

Note*: γ denotes the Euler-Mascheroni constant.*

are generated. An ensemble of 50 members is drawn randomly from each of the forecast distributions, which is then paired with the observations.

The performance of the five forecasters is evaluated using the absolute error, the squared error, the ignorance score, the CRPS, and the PIT histogram. We also produced rank histograms, but they turned out to be almost identical to the PIT histograms. As we encountered variations in the scores depending on the initial random seed, the whole process is repeated 10 times with different initial seeds, so that the final number of forecast-observation pairs comes to 1000.

In order to understand the true ranking of the five forecasting methods in terms of skill, we reproduce the simulation study with 10 times 100,000 forecasts. For the case of a normal true distribution, Fig. 6.3 shows the mean absolute error, mean CRPS and mean ignorance score, along with 95% bootstrap confidence intervals (see Section 6.3.7) computed from 1000 bootstrap samples. We have omitted the squared error from this plot, as its values are on a much larger scale than the other scores. Looking at the results for the small sample size in the top row, all scores assign the lowest mean value, and therefore the highest skill, to the normal distribution with the true parameters. However, if no knowledge about the true distribution is available, as in a real forecast setting, the absolute error and the CRPS

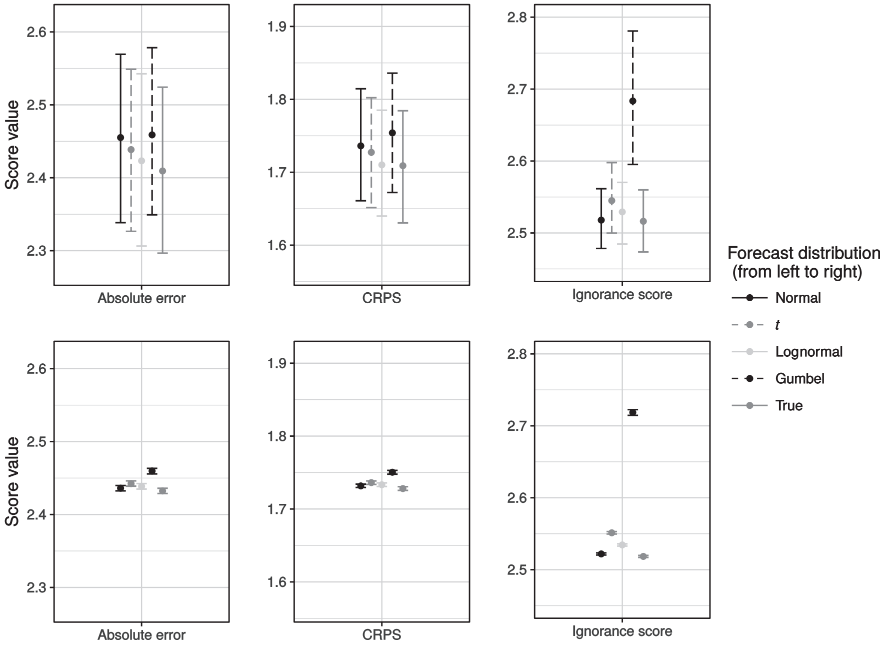


FIG. 6.3

Top row: Mean absolute error, CRPS and ignorance score, and the 95 % bootstrap confidence interval for the five forecast distributions, if the true distribution is normal. Scores are based on 1000 forecast-observation pairs. *Bottom row*: Same as above, but scores are based on 1 million forecast-observation pairs.

would prefer the lognormal distribution over all other forecasters, while the ignorance score judges the normal distribution with estimated parameters to be the best.

The bottom panel of Fig. 6.3 shows the results from running the same study with the larger sample size, which changes the order in which we would expect the forecasters to rank. Here, all scores correctly find the Gumbel distribution, which has a completely different shape and tail behavior than the truth, to be the worst forecast, and the two forecasts based on normal distributions to be the best. This contradicts the results in the top panel, where only the ignorance score ranked the forecasters in the same order as we would expect.

Due to assigning large penalties to outliers, the ignorance score is able to discriminate between the shapes of the forecast distributions, and shows a significant difference at the 95% level between the Gumbel and the normal, lognormal, and true distributions. The relatively poor performance of the non-central t-distribution can probably be explained by the fact that, while this distribution approximates a normal distribution if the degrees of freedom are large, the asymptotic distribution will have a standard deviation of 1, which does not match the given standard deviation of 3 in this example.

Judging from Fig. 6.4, which shows PIT histograms for the small-sample study with a normal true distribution, we cannot make any statements about the forecast ranking, except that the Gumbel distribution forecast is clearly uncalibrated. Only when looking at the large sample equivalent in Fig. 6.5 do we see that the normal and the true forecasters are the only ones not suffering from miscalibration. A formal chi-squared test (see Section 6.3.7) rejects the assumption of uniformity for the Gumbel

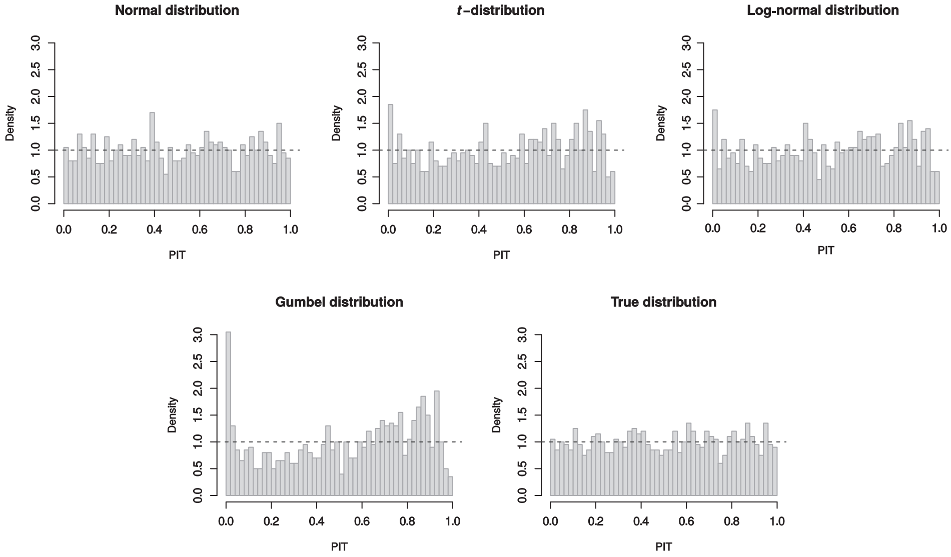


FIG. 6.4

PIT histograms for the five forecast distributions, if the true distribution is normal, based on 1000 forecast-observation pairs.

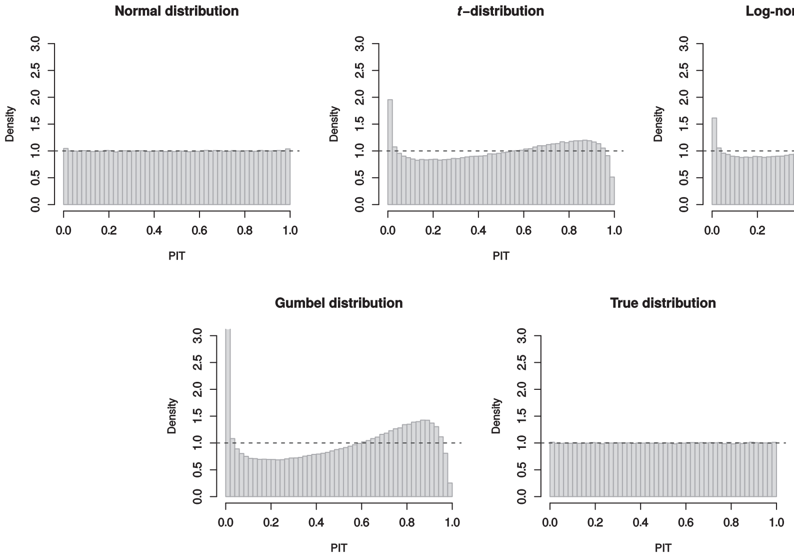


FIG. 6.5

PIT histograms for the five forecast distributions, if the true distribution is normal, based on 1 million forecast-observation pairs.

distribution and even the *t* and lognormal distributions (at a level of 5%) in the small-sample case, and for all distributions apart from the true one in the large sample case.

Fig. 6.6 illustrates one example forecast, for which the scores are plotted as functions of the verifying observation, in this case a sample value from a $\mathcal{N}(27.16, 9)$ distribution. While the score minima largely coincide for the true and the *t*-distribution, it becomes clear from the shape of the ignorance score why it is much better at identifying the Gumbel distribution as inferior: because of the lack of symmetry, Gumbel forecasts will receive a much higher penalty if the observation lies left of the distribution mode than if it lies on the right.

For the second part of the simulation study, we used a Gumbel distribution as truth, where the mean is distributed as $\mathcal{N}(25, 1)$ and the scale parameter is 3. The same kinds of forecasts are produced again: normal, noncentral *t*, lognormal, and Gumbel distributions, based on the sample means and variances of the training data. In Fig. 6.7, the outcome of the study is shown for a small sample size (top row) and a very large sample size (bottom row). As previously, all scores agree on the forecast ranking when the sample is large. The Gumbel distribution with estimated parameters and the true Gumbel distribution are assigned the lowest scores, while the normal forecaster now has the lowest skill.

However, the rankings look different in the top panel, where the true distribution is only ranked the third best by the absolute error and the CRPS, behind the estimated Gumbel and noncentral *t*-distributions. The ignorance score again is the only score able to reproduce the forecast ranking we expect from the bottom panel. This is, of course, concerning and hints at the fact that even for a data set of apparently sufficient size, such as the 1000 50-member ensembles used here, the scores do not necessarily provide robust and proper results.

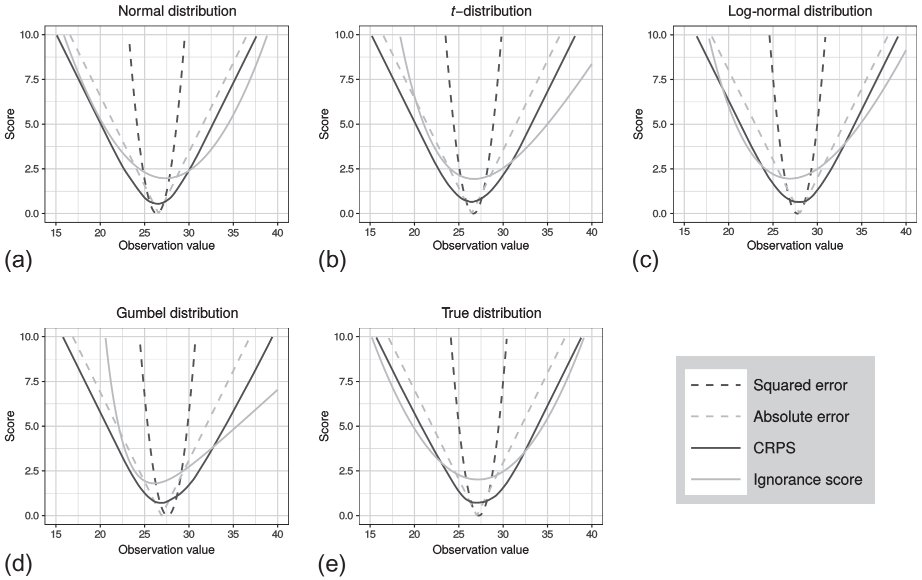


FIG. 6.6

Squared error, absolute error, CRPS, and ignorance score as functions of the verifying observation, for one forecast case in the simulation study: (a) normal distribution forecast, (b) noncentral *t*-distribution forecast, (c) lognormal distribution forecast, (d) Gumbel distribution forecast, and (e) forecast based on the true normal distribution.

Again we cannot really judge the degree of forecast calibration by just looking at the small-sample PIT histograms in Fig. 6.8, except for the clearly uncalibrated normal distribution. A case could be made that the histogram for the true distribution looks slightly flatter than the other ones, but not with great certainty. It becomes clear, however, from Fig. 6.9, that the forecasts based on noncentral *t* and lognormal distributions also suffer from multiple types of miscalibration. These findings are confirmed by a chi-squared test, which rejects the uniformity hypothesis for all except the Gumbel distributions in Fig. 6.8 and all except the true distribution in Fig. 6.9.

Picking an example forecast from the data set, Fig. 6.10 shows that the ignorance score for the two Gumbel distribution forecasters is again nonsymmetric, and therefore minimizes at a different value compared with the CRPS. In general, the ignorance score takes its minimum value at the mode of the distribution, and the CRPS at the median.

We can gather from this simulation study that even proper scores can behave very differently, depending on the size of the underlying data set, and are not necessarily able to rank competing forecasters according to their actual skill. Therefore, we suggest always using a combination of scoring rules to get a maximum amount of information about the performance of a particular model or forecaster. The ignorance score is more sensitive to the shape of a distribution and thus is suitable to check if a chosen distribution actually fits the data. The CRPS is very useful for comparing models when the forecasts do not take the form of a standard probability distribution, or if for a given data set such a distribution cannot be perfectly specified.

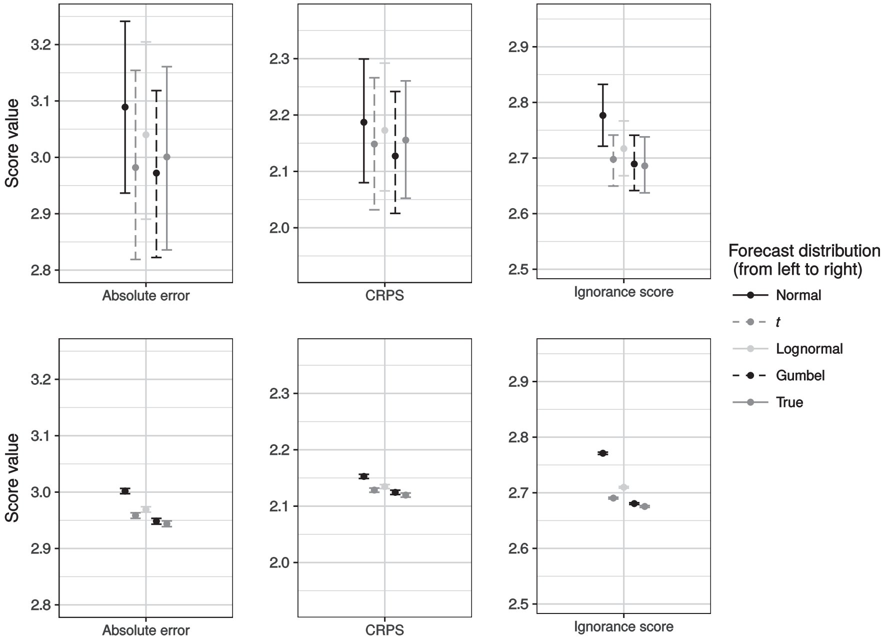


FIG. 6.7

Top row: Mean absolute error, CRPS and ignorance score, and the 95 % bootstrap confidence interval for the five forecast distributions, if the true distribution is a Gumbel distribution. Scores are based on 1000 forecast-observation pairs. *Bottom row*: Same as above, but scores are based on 1 million forecast-observation pairs.

These results also have implications for the ongoing discussion of whether to use maximum likelihood methods or minimize the CRPS to estimate model parameters (Gneiting, Raftery, Westveld, & Goldman, 2005), in that there might not be a definitive answer. Depending on the forecast situation and model choice, it could be preferable to switch between the two approaches. A case can be made for performing a thorough exploratory analysis of the data at hand before fitting any distributions, to find one that matches the data best. If it is difficult to select one distribution over the other, the simpler model should be preferred.

In all circumstances, the ranking of forecasters should not be solely based on the mean score, even if the sample size seems to be sufficiently large, but confidence intervals should be given, for example, by applying bootstrapping techniques. We found that even for 1 million data points, differences between the forecast scores were often not significant at the 5% level.

6.3.3 ASSESSING EXTREME EVENTS

Forecasts specifically aimed at predicting extreme events can be assessed in a standard manner, for example, by using the scoring rules discussed in Section 6.3.1 (Friederichs & Thorarinsdottir, 2012).

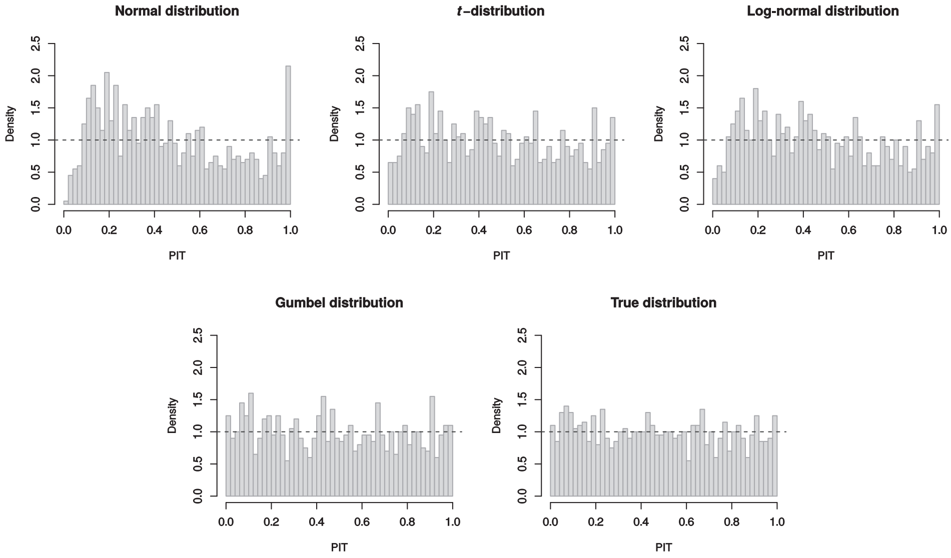


FIG. 6.8

PIT histograms for the five forecast distributions, if the true distribution is a Gumbel distribution, based on 1000 forecast-observation pairs.

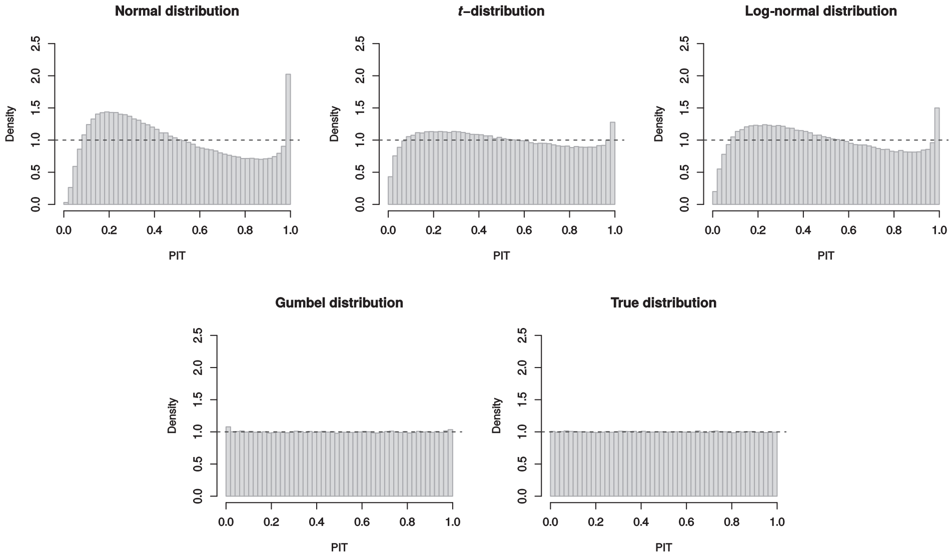


FIG. 6.9

PIT histograms for the five forecast distributions, if the true distribution is a Gumbel distribution, based on 1 million forecast-observation pairs.

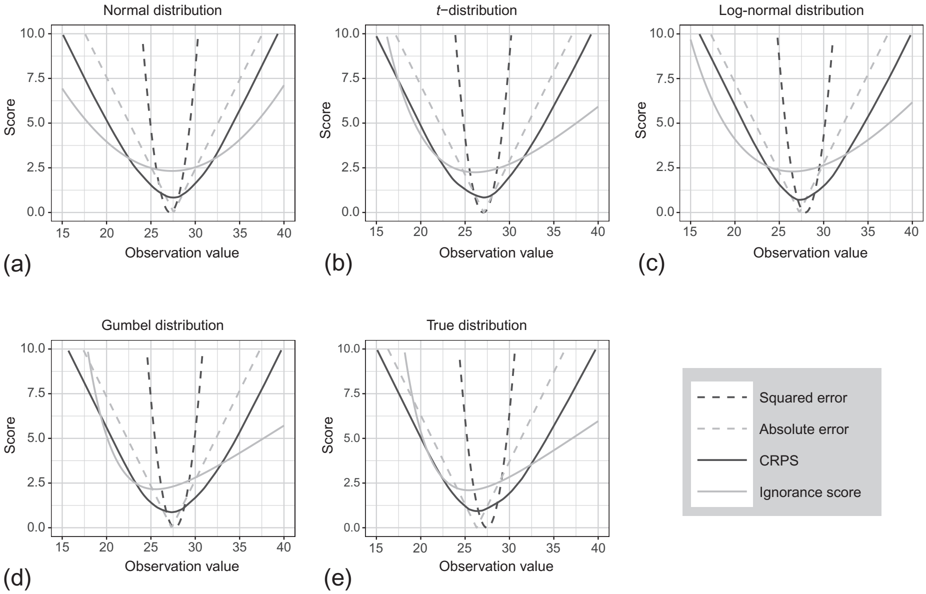


FIG. 6.10

Squared error, absolute error, CRPS, and ignorance score as functions of the verifying observation, for one forecast case in the simulation study: (a) normal distribution forecast, (b) noncentral *t*-distribution forecast, (c) lognormal distribution forecast, (d) Gumbel distribution forecast, and (e) forecast based on the true Gumbel distribution.

However, the restriction of conventional forecast evaluation to subsets of extreme observations by selecting the extreme observations after-the-fact while discarding the nonextreme ones, and to proceed with standard evaluation tools, will invalidate their theoretical properties and encourage hedging strategies (Lerch, Thorarinsdottir, Ravazzolo, & Gneiting, 2017).

Specifically, Gneiting and Ranjan (2011) show that a proper scoring rule S is rendered improper if the product with a nonconstant weight function w is formed, where w depends on the observed value y. That is, consider the weighted scoring rule

$$S_0(F, y) = w(y)S(F, y). \tag{6.22}$$

Then if Y has density g, the expected score $\mathbb{E}_g S_0(F, Y)$ is minimized by the predictive distribution F with density

$$f(y) = \frac{w(y)g(y)}{\int w(z)g(z)\mathrm{d}z} \tag{6.23}$$

which is proportional to the product of the weight function w and the true density g. In particular, if $w(y) = \mathbb{1}\{y \geq u\}$ for some high threshold value u, then S_0 corresponds to evaluating F only on observed values exceeding u under the scoring rule S.

Instead, one can apply proper *weighted scoring rules* that are tailored to emphasize specific regions of interest. Diks, Panchenko, and Van Dijk (2011) propose two weighted versions of the ignorance score that correct for the result in Eq. (6.23). The *conditional likelihood* (CL) score is given by

$$\mathrm{CL}(F,y) = -w(y)\log\left(\frac{f(y)}{\int_{\Omega} w(z)f(z)\mathrm{d}z}\right)$$

and the *censored likelihood* (CSL) score is defined as

$$\mathrm{CSL}(F,y) = -w(y)\log f(y) - (1-w(y))\log\left(1-\int_{\Omega} w(z)f(z)\mathrm{d}z\right)$$

Here, w is a weight function such that $0 \le w(y) \le 1$ and $\int w(y)f(y)\mathrm{d}y > 0$ for all potential predictive distributions $F \in \mathcal{F}$. When $w(y) \equiv 1$, both the CL and the CSL score reduce to the unweighted ignorance score in Eq. (6.9).

Gneiting and Ranjan (2011) propose the *threshold-weighted continuous ranked probability score* (twCRPS), defined as

$$\mathrm{twCRPS}(F,y) = \int_{\Omega} w(z)(F(z) - \mathbb{1}\{y \le z\})^2 \mathrm{d}z$$

where, again, w is a nonnegative weight function, see also Matheson and Winkler (1976). When $w(y) \equiv 1$, the twCRPS reduces to the unweighted CRPS in Eq. (6.12) while $w(y) = \mathbb{1}\{y = u\}$ equals the Brier score in Eq. (6.14). More generally, the twCRPS puts emphasis on a particular part of the forecast distribution F as specified by w. For focusing on the upper tail of F, Gneiting and Ranjan (2011) consider both indicator weight functions of the type $w(y) = \mathbb{1}\{y \ge u\}$ and nonvanishing weight functions such as $w(y) = \Phi(y|u, \sigma^2)$ where Φ denotes the cumulative distribution function of the Gaussian distribution with mean u and variance σ^2. Corresponding weight functions for the lower tail of F are given by $w(y) = \mathbb{1}\{y \le u\}$ and $w(y) = 1 - \Phi(y|u, \sigma^2)$ for some low threshold value u.

Nonstationarity in the mean climate, for example, due to spatial heterogeneity, may render it difficult to define a common threshold value u over a large number of forecast cases. Here, it may be more natural to define a weight function in quantile space using the CRPS representation in Eq. (6.13),

$$\mathrm{twCRPS}(F,y) = \int_0^1 w(\tau)\left(F^{-1}(\tau) - y\right)\left(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau\right)\mathrm{d}\tau$$

where w is a nonnegative weight function on the unit interval (Gneiting & Ranjan, 2011; Matheson & Winkler, 1976). Setting $w(\tau) \equiv 1$ retrieves the unweighted CRPS in Eq. (6.13) while this definition of twCRPS with $w(\tau) = \mathbb{1}\{\tau = q\}$ equals the quantile score in Eq. (6.15). Examples of more general weight functions for this setting include $w(\tau) = \mathbb{1}\{\tau \ge q\}$ and $w(\tau) = \tau^2$ for the upper tail, and $w(\tau) = \mathbb{1}\{\tau \le q\}$ and $w(\tau) = (1-\tau)^2$ for the lower tail, with appropriate threshold values q, see also Gneiting and Ranjan (2011).

Lerch et al. (2017) find that there are limited benefits in using weighted scoring rules compared with using standard, unweighted scoring rules when testing for equal predictive performance. However, the application of weight functions as described here may facilitate interpretation of the forecast skill.

6.3.4 EXAMPLE: PROPER AND NONPROPER VERIFICATION OF EXTREMES

In the following, we illustrate that the use of nonproper methods to verify and compare competing forecasts for extremes can lead to a distortion of the results and possibly false inference. Taking the same setting as the first part of the simulation study in Section 6.3.2, we generate sets of observation and training data from a normal distribution with standard deviation 3 and the mean a random value from a $\mathcal{N}(25,1)$ distribution.

Four of the forecasting methods in Section 6.3.2 are compared: a normal distribution with estimated parameters based on the training data, a Gumbel distribution with estimated parameters, a normal distribution with the true parameters, and a Gumbel distribution with the true means as location parameter and scale parameter $\sigma=3$. The forecasters' performance for extremes, which we consider to be values greater or equal to the 97.5% quantile of the observations u, will be measured using the threshold-weighted CRPS with three different weight functions and the unweighted CRPS, where the cases are restricted to observations above the threshold. The weight functions considered are variations on the indicator function:

$$\begin{aligned} w_1(y) &= \mathbb{1}\{y \geq u\} \\ w_2(y) &= 1 + \mathbb{1}\{y \geq u\} \\ w_3(y) &= 1 + \mathbb{1}\{y \geq u\} \cdot u \end{aligned}$$

Mean scores and 95% confidence intervals, calculated by numerical integration based on the small sample data set from Section 6.3.2, are shown in Fig. 6.11 for the threshold-weighted CRPS and the CRPS with restricted observations, along with the unweighted CRPS. The results for the twCRPS with weight function w_1 are omitted, as they are equal to 0 for all forecasters.

However, just by adding 1 to the indicator function, we obtain meaningful scores with weight function w_2, showing the Gumbel distribution with fixed parameters to be the least skillful forecast, while the two normal distribution forecasters are of significantly better quality. The twCRPS with weight function w_3 and the unweighted CRPS lead to similar conclusions, although the differences between the scores are sometimes not significant. In contrast to the other scores, the CRPS based on the restricted data set clearly shows the Gumbel distribution with fixed parameters to be the preferred forecaster.

Although the fixed Gumbel parameters and shape are obviously wrong, this is no surprise, as this distribution was purposely chosen because it has a heavy tail. Fig. 6.12 shows predictive densities for one example from the data set. If we restrict the evaluation to the area above the chosen threshold, represented by the black vertical line, the Gumbel distribution with fixed parameters is indeed the seemingly best forecast, as it assigns the highest probabilities to extreme values. The two normal distributions and the Gumbel distribution with estimated parameters, which tries to approximate the true normal distribution, have a very similar tail behavior, explaining their similar performance in terms of all scores.

We come to the same conclusion as Lerch et al. (2017), that conditioning a data set on extremal observations can result in preferring a forecaster who predicts extremes with inflated probabilities. When evaluating forecasts for a certain range of values, proper methods such as the threshold-weighted CRPS should be used, where the whole data set is considered.

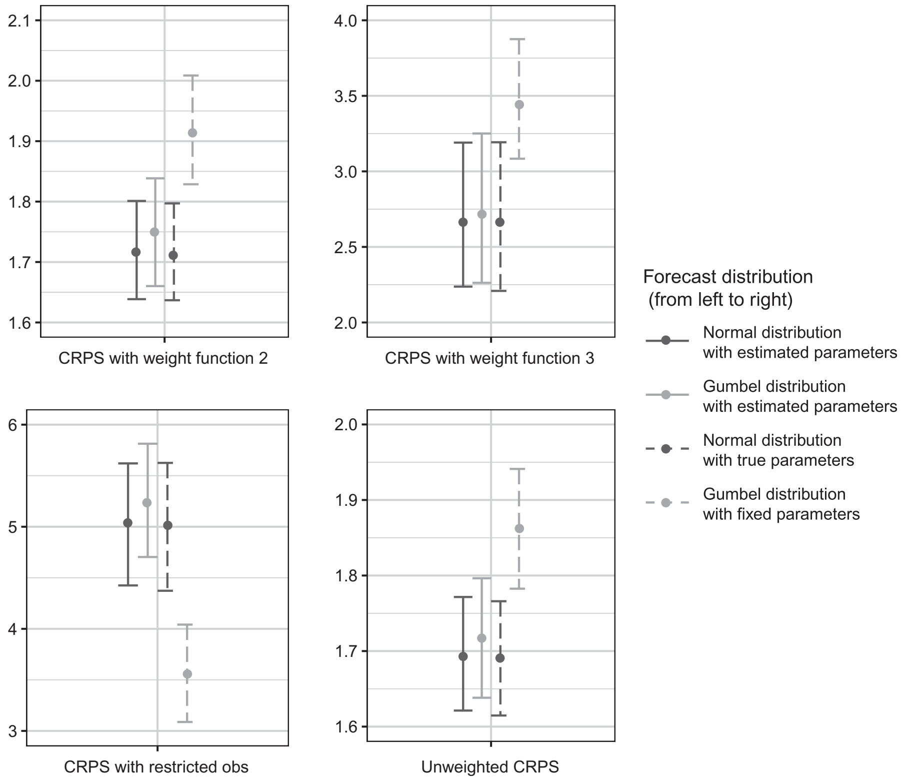


FIG. 6.11

Mean scores and 95% bootstrap confidence interval for the four versions of the CRPS. *Top row*: twCRPS with weight functions w_2 and w_3. *Bottom row*: CRPS restricted to observations above the threshold u and unweighted CRPS.

6.3.5 MULTIVARIATE ASSESSMENT

Two general approaches can be employed to assess multivariate forecasts with scoring rules: Use specialized multivariate scores, or reduce the multivariate forecast to a univariate quantity and subsequently apply the univariate scores discussed previously. For the latter approach, the appropriate univariate quantities depend on the context. Multivariate forecasts of single weather quantities are usually in the form of temporal trajectories, spatial fields, or space-time fields. Here it can, for instance, be useful to assess the predictive performance of derived quantities such as maxima, minima, and accumulated totals, all of which depend on accurate modeling of both marginal and higher order structures. See, for example, Feldmann, Scheuerer, and Thorarinsdottir (2015) for an assessment of spatial forecast fields for temperature.

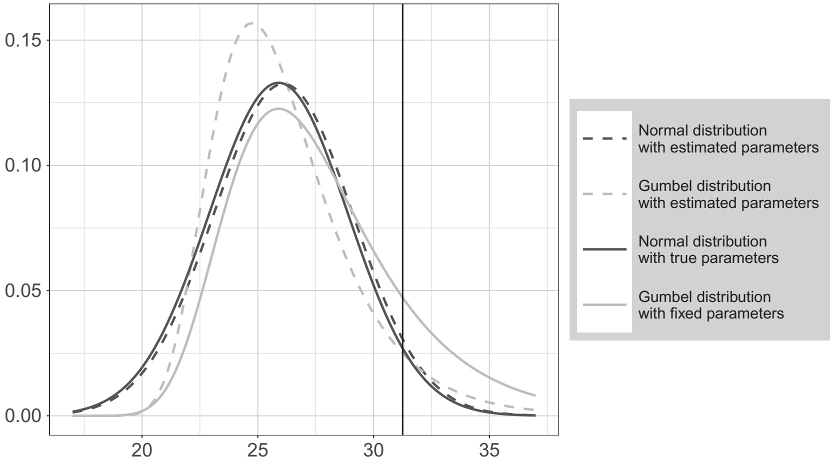


FIG. 6.12

Example predictive densities given by the four competing forecasters. The *black vertical line* shows the threshold *u*, above which observations are considered to be extreme.

Scores that directly assess multivariate forecasts are rather scarce and, as noted by Gneiting and Katzfuss (2014), there is a need to develop further decision-theoretically principled methods for multivariate assessment. The univariate Dawid-Sebastiani score in Eq. (6.10) can be applied in a multivariate setting with

$$\mathrm{DS}(F,y) = \log \det \Sigma_F + (y-\mu_F)^{\top} \Sigma_F^{-1} (y-\mu_F) \tag{6.24}$$

where μ_F is the mean vector and Σ_F the covariance matrix of the predictive distribution with $\det \Sigma_F$ denoting the determinant of Σ_F (Dawid & Sebastiani, 1999). However, note that unless the sample size is much larger than the dimension of the multivariate quantity, sampling errors can affect the calculation of $\det \Sigma_F$ and Σ_F^{-1} (see e.g., Table 2 in Feldmann et al., 2015). Similarly, if the multivariate predictive density is available, the ignorance score in Eq. (6.9) can be employed (Roulston & Smith, 2002).

Gneiting and Raftery (2007) propose the *energy score* (ES) as a multivariate generalization of the CRPS. It is given by

$$\mathrm{ES}(F,y) = \mathbb{E}_F \parallel X-y \parallel - \frac{1}{2}\mathbb{E}_F\mathbb{E}_F \parallel X-X' \parallel \tag{6.25}$$

where X and X' are two independent random vectors distributed according to F and $\|\cdot\|$ is the Euclidean norm. For ensemble forecasts, the natural analog of the formulas in Eqs. (6.16), (6.17) apply. If the multivariate observation space Ω^d consists of weather variables on varying scales, the margins should be standardized before computing the joint energy score for these variables (Schefzik, Thorarinsdottir, & Gneiting, 2013). This can be done using the marginal means and standard deviations of the observations in the test set. The energy score has been developed with low-dimensional quantities in mind and it may lose discriminatory power in higher dimensions (Pinson, 2013).

Scheuerer and Hamill (2015) propose a multivariate scoring rule that considers pairwise differences of the components of the multivariate quantity. In its general form, the *variogram score (VS) of order p* is given by

$$\mathrm{VS}_p(F,y) = \sum_{i=1}^{d}\sum_{j=1}^{d}\omega_{ij}\left(|y_i - y_j|^p - \mathbb{E}_F|X_i - X_j|^p\right)^2 \tag{6.26}$$

where y_i and y_j are the ith and the jth component of the observation, X_i and X_j are the ith and the jth component of a random vector X that is distributed according to F, and ω_{ij} are nonnegative weights. Scheuerer and Hamill (2015) compare different choices of the order p and find that the best results in terms of discriminative power are obtained with $p = 0.5$. Furthermore, they recommend using weights proportional to the inverse distance between the components unless a prior knowledge regarding the correlation structure is available.

A comparison of the three multivariate scores in Eqs. (6.24)–(6.26) is provided in Scheuerer and Hamill (2015). The authors conclude by recommending the use of multiple scores as they complement each other in their strengths and weaknesses. The variogram score is generally able to distinguish between correct and misspecified correlation structures, but it has certain limitations resulting from the fact that it is proper but not strictly proper. Some of these limitations can be addressed by also using the energy score that is more sensitive to misspecifications in the predictive mean and less affected by finite representations of the predictive distribution. While the latter is an issue for the Dawid-Sebastiani score, it performs well for continuous predictive distributions, in particular for multivariate Gaussian models (Wei, Balabdaoui, & Held, 2017).

6.3.6 DIVERGENCE FUNCTIONS

In some cases, in particular in climate modeling, it is of interest to compare the predictive distribution F against the true distribution of the observations, which is commonly approximated by the *empirical distribution function* of the available observations $y_1, \ldots, y_n$,

$$\hat{G}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{y_i \le x\}. \tag{6.27}$$

The two distributions, F and $\hat{G}_n$, can be compared using a *divergence*

$$D:\ \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\ge 0} \tag{6.28}$$

where $D(F, F) = 0$.

Assume that the observations $y_1, \ldots, y_n$ forming the empirical distribution function $\hat{G}_n$ are independent with distribution $G \in \mathcal{F}$. A propriety condition for divergences corresponding to that for scoring rules (Eq. 6.7) states that the divergence D is *n-proper* for a positive integer n if

$$\mathbb{E}_G D(G, \hat{G}_n) \le \mathbb{E}_G D(F, \hat{G}_n) \tag{6.29}$$

and *asymptotically proper* if

$$\lim_{n\to\infty}\mathbb{E}_G D(G, \hat{G}_n) \le \lim_{n\to\infty}\mathbb{E}_G D(F, \hat{G}_n) \tag{6.30}$$

for all probability distributions $F, G \in \mathcal{F}$ (Thorarinsdottir, Gneiting, & Gissibl, 2013). While the condition in Eq. (6.30) is fulfilled by a large class of divergences, only score divergences have been shown to fulfill Eq. (6.29) for all integers n. A divergence D is a *score divergence* if there exists a proper scoring rule S such that $D(F,G) = \mathbb{E}_G S(F,Y) - \mathbb{E}_G S(G,Y)$.

A score divergence that assesses the full distributions is the *integrated quadratic divergence* (IQD)

$$\mathrm{IQD}(F,G) = \int_{-\infty}^{+\infty} (F(x) - G(x))^2 \mathrm{d}x \tag{6.31}$$

which is the score divergence of the continuous ranked probability score (Eq. 6.12). Alternative score divergences that assess specific properties of the predictive distribution include the *mean value divergence* (MVD),

$$\mathrm{MVD}(F,G) = (\mathrm{mean}(F) - \mathrm{mean}(G))^2 \tag{6.32}$$

which is the divergence associated with the squared error scoring rule (Eq. 6.20), and the *Brier divergence* (BD) associated with the Brier score (Eq. 6.14),

$$\mathrm{BD}(F,G|u) = (G(u) - F(u))^2 \tag{6.33}$$

for some threshold u.

Fig. 6.13 provides a comparison of the score divergences in Eqs. (6.31)–(6.33) for two simple settings where the observation distribution is given by a standard normal distribution and all the forecast distributions are also normal distributions but with varying parameters. In the left plot, the variance is correctly specified while the forecast mean value varies. In the right plot, the forecast mean values equal that of the observation distribution while the standard deviation varies. We compare the IQD, the MVD, and the BD with thresholds $u = 0.67$ and $u = 1.64$, which equal the 75% and the 95% quantiles of the observation distribution, respectively. The divergences are more sensitive to forecast errors in the mean than the spread. In particular, the MVD is, naturally, not able to detect errors in the forecast spread. Furthermore, integrating over the BD for all possible thresholds u and obtaining the IQD yields

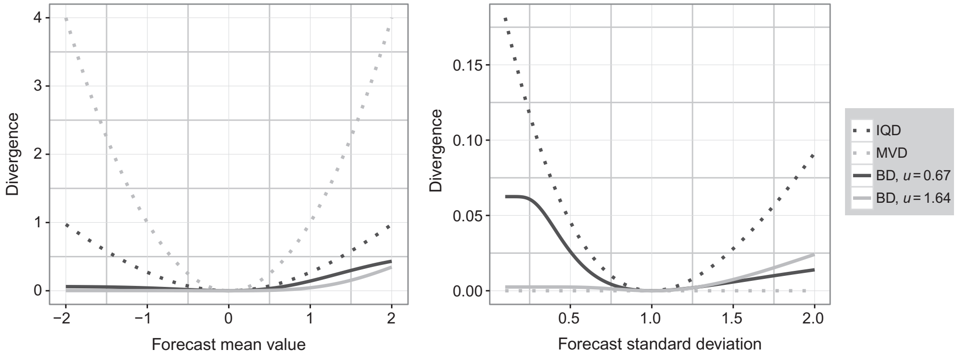


FIG. 6.13

Comparison of expected score divergence values for a standard normal observation distribution and normal forecast distributions with varying mean values (*left*) or standard deviations (*right*).

a better discrimination than investigating the differences for individual quantiles. The right plot also shows that the model ranking obtained under the BD strongly depends on the threshold u.

While every proper scoring rule is associated with a score divergence, not all score divergences are practical for use in the setting where the empirical distribution function $\hat{G}_n$ is used. One example is the Kullback-Leibler divergence, which is the score divergence of the ignorance score in Eq. (6.9). The Kullback-Leibler divergence becomes ill-defined if the forecast distribution F has positive mass anywhere where the observation distribution G has mass zero. When G is replaced by $\hat{G}_n$ and, especially, if the sample size n is relatively small, such issues might occur. One option to circumvent the issue is to treat the data as categorical and bin it in b bins prior to the evaluation. That is, identify the probability distribution F with a probability vector $(f_1, \ldots, f_b)$ and, similarly, G with a probability vector $(g_1, \ldots, g_b)$. The *Kullback-Leibler divergence* is then given by

$$\mathrm{KLD}(F,G) = \sum_{i=1}^{b} f_i \log \frac{f_i}{g_i}$$

see also the discussion in Thorarinsdottir et al. (2013).

Historically, much of the forecast evaluation literature has focused on the evaluation of probabilistic forecasts against deterministic observations and an in-depth discussion of optimal theoretical and/or practical properties of divergences is lacking. Applied studies commonly employ divergences that are asymptotically proper rather than n-proper for all positive integer n, see for example, Palmer (2012) and Perkins, Pitman, Holbrook, and McAneney (2007).

6.3.7 TESTING EQUAL PREDICTIVE PERFORMANCE

As demonstrated in the simulation study in Section 6.3.2, the estimation of the mean score over a test set may be associated with a large uncertainty. A simple bootstrapping procedure over the individual scores may be used to assess the uncertainty in the mean score, see for example, Friederichs and Thorarinsdottir (2012). Assume we have n score values $S(F_1, y_1), \ldots, S(F_n, y_n)$. By repeatedly resampling vectors of length n (with replacement) and calculating the mean of each sample, we obtain an estimate of the variability in the mean score. Note that some care is needed if the forecast errors, and thus the resulting scores, are correlated. A comprehensive overview over bootstrapping methods for dependent data is given in Lahiri (2003).

Formal statistical tests can be applied to test equal predictive performance of two competing methods under a proper scoring rule. The most commonly applied test is the *Diebold-Mariano test* (Diebold & Mariano, 1995), which applies in the time series setting. Consider two competing forecasting methods F and G that for each time step $t = 1, \ldots, n$ issue forecasts F_t and G_t, respectively, for an observation y_{t+k} that lies k time steps ahead. The mean scores under a scoring rule S are given by

$$\overline{S}_n^F = \frac{1}{n}\sum_{t=1}^{n} S(F_t, y_{t+k}) \quad \text{and} \quad \overline{S}_n^G = \frac{1}{n}\sum_{t=1}^{n} S(G_t, y_{t+k})$$

The Diebold-Mariano test uses the test statistic

$$t_n = \sqrt{n}\,\frac{\overline{S}_n^F - \overline{S}_n^G}{\hat{\sigma}_n} \tag{6.34}$$

where $\hat{\sigma}_n^2$ is an estimator of the asymptotic variance of the score difference. Under the null hypothesis of equal predictive performance and standard regularity conditions, the test statistic t_n in Eq. (6.34) is asymptotically standard normal (Diebold & Mariano, 1995). When the null hypothesis is rejected in a two-sided test, F is preferred if t_n is negative and G is preferred if t_n is positive.

Diebold and Mariano (1995) note that for ideal k-step-ahead forecasts, the forecast errors are at most $(k-1)$-dependent. An estimator for the asymptotic variance $\hat{\sigma}_n^2$ based on this assumption is given by

$$\hat{\sigma}_n^2 = \begin{cases} \hat{\gamma}_0 & \text{if } k = 1 \\ \hat{\gamma}_0 + 2\sum_{j=1}^{k-1} \hat{\gamma}_j, & \text{if } k \geq 2 \end{cases} \tag{6.35}$$

where $\hat{\gamma}_j$ denotes the lag j sample autocorrelation of the sequence $\{S(F_i, y_{i+k}) - S(G_i, y_{i+k})\}_{i=1}^n$ for $j = 0, 1, 2, \ldots$ (Gneiting & Ranjan, 2011). Alternative estimators are discussed in Diks et al. (2011) and Lerch et al. (2017).

In the spatial setting, Hering and Genton (2011) propose the *spatial prediction comparison test*, which accounts for spatial correlation in the score values without imposing assumptions on the underlying data or the resulting score differential field. This test is implemented in the `R` package `SpatialVx` (Gilleland, 2017). Weighted scoring rules and their connection to hypothesis testing are discussed in Holzmann and Klar (2017).

A simple test for the uniformity of a rank or PIT histogram is the chi-squared test. It tests if the histogram values can be considered samples from a uniform distribution and therefore if any deviations of uniformity are random or systematic (Wilks, 2004, 2011). The chi-squared statistic based on n cases and K ensemble members is

$$\chi^2 = \sum_{i=1}^{K+1} \frac{(m_i - f)^2}{f} \tag{6.36}$$

with m_i denoting the actual number of counts for bin i and $f = \frac{n}{K+1}$ the expected number of counts for a uniform distribution. We can reject the null hypothesis of the histogram being uniform if this statistic exceeds the quantile of the chi-squared distribution with K degrees of freedom at the chosen level of significance.

In its general form, however, the chi-squared test only applies to independent data, which is not the case in many forecast settings due to, for example, temporal or spatial correlation between forecast data points. Some methods to address this effect are proposed in Wilks (2004). If the goal is to not only test for uniformity, but also for the other deficiencies in calibration shown in Section 6.2.1, Elmore (2005) and Jolliffe and Primo (2008) present alternatives that are more flexible and appropriate. Wei et al. (2017) propose calibration tests for multivariate Gaussian forecasts based on the Dawid-Sebastiani score in Eq. (6.24).

6.4 UNDERSTANDING MODEL PERFORMANCE

When assessing the performance of an individual model, for example, to identify weaknesses and test potential improvements, it might be useful to look at tools that do not necessarily follow the principles of propriety described in Section 6.3. For instance, it can be useful to investigate the forecast bias to better understand the potential sources of forecast errors even if competing forecasting models should not be ranked based on mean bias as it is not a proper score (Gneiting & Raftery, 2007). Here, we discuss a few tools that may be used to provide a better understanding of the performance of an individual forecasting model, even though ranking of competing forecasters should not be based on these tools.

One of the most popular measures used by national weather services is the anomaly correlation coefficient (ACC), a valuable tool to track the gain in forecast skill over time (Jolliffe & Stephenson, 2012). The ACC quantifies the correlation between forecast anomalies and the anomalies of the observation, typically an analysis. Anomalies are defined as the difference between the forecast or analysis and the climatology for a given time and location. Usually, the climatology is based on the model climate, calculated from the range of values predicted by the dynamical forecast model over a long time period.

For a deterministic forecast f_i, valid at time i, with a corresponding analysis a_i and climate statistic c_i, there are two equivalent definitions for the ACC (e.g., Miyakoda, Hembree, Strickler, & Shulman, 1972):

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N}(f_i - c_i)\cdot(a_i - c_i) - \sum_{i=1}^{N}(f_i - c_i)\cdot\sum_{i=1}^{N}(a_i - c_i)}{\sqrt{\sum_{i=1}^{N}(f_i - c_i)^2 - \left(\sum_{i=1}^{N}(f_i - c_i)\right)^2}\cdot\sqrt{\sum_{i=1}^{N}(a_i - c_i)^2 - \left(\sum_{i=1}^{N}(a_i - c_i)\right)^2}}$$

$$= \frac{\sum_{i=1}^{N}\left(f'_i - \overline{f}'\right)\left(a'_i - \overline{a}'\right)}{\sqrt{\sum_{i=1}^{N}\left(f'_i - \overline{f}'\right)^2\sum_{i=1}^{N}\left(a'_i - \overline{a}'\right)^2}}$$

Here, $f'_i = f_i - c_i$ is the forecast anomaly and $a'_i = a_i - c_i$ the anomaly of the analysis, with respective sums $\overline{f}' = \sum_{i=1}^{N}(f_i - c_i)$ and $\overline{a}' = \sum_{i=1}^{N}(a_i - c_i)$. The ACC is a preferred evaluation tool for gridded forecasts and spatial fields, as these are usually compared with an analysis or a similar gridded observation product.

However, there are certain limitations and pitfalls one has to be aware of when using this measure. Due to it being a correlation coefficient, the ACC does not give any information about forecast biases and errors in scale, so that it can overestimate the forecast skill (Murphy & Epstein, 1989). As such, it should always be used in conjunction with an estimate of the actual bias, or applied to previously bias-corrected data.

It has been established empirically that an anomaly correlation of 0.6 corresponds to a limit in usefulness for a medium-range forecast. Murphy and Epstein (1989) warn, however, that the ACC is an upper limit of the actual skill and that the ACC should be seen as a measure of potential skill. Naturally, the ACC relies to a large extent on the underlying climatology used to compute the anomalies.

When evaluating forecast skill with proper scores, it is often useful to compute separate indicators for the degree of calibration and the sharpness of the forecast. The well-known and widely used decomposition of the Brier score by Murphy (1973b) separates the score value in three parts, quantifying reliability, resolution, and uncertainty.

Consider a forecast sample of size N, where probability forecasts $p_u = 1 - F(u)$ are computed for exceeding a threshold u and binary observations take the form $o = \mathbb{1}\{y \geq u\}$. If the forecasts take K unique values, with n_k denoting the number of forecasts within the category k and $p_{u,k}$ the probability forecast associated with category k, then the Brier score can be written as

$$\mathrm{BS}(F,y|u) = \frac{1}{N}\sum_{k=1}^{K} n_k (p_{u,k} - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o}) \tag{6.37}$$

where $\bar{o}_k$ is the event frequency for each of the forecast values and $\bar{o} = \frac{1}{N}\sum_{i=1}^{N} o_i$ the climatological event frequency, computed from the sample. The first part of the sum in Eq. (6.37) relates to the reliability or calibration, the second, having a negative effect on the total score, to the resolution or sharpness, and the last part is the climatological uncertainty of the event.

This representation of the Brier score relies on the number of discrete forecast values K being relatively small. If p_u takes continuous values, care must be taken when binning the forecast into categories, so as not to introduce biases (Bröcker, 2008; Stephenson, Coelho, & Jolliffe, 2008). Several analog decompositions have been proposed for other scores, such as the CRPS (Hersbach, 2000), the quantile score (Bentzien & Friederichs, 2014), and the ignorance score (Weijs, van Nooijen, & van de Giesen, 2010). Bröcker (2009) shows that any proper score can be decomposed analogously to Eq. (6.37). Recently, Siegert (2017) formulated a general framework allowing for the decomposition of arbitrary scores.

While it is common and advisable to look at a model's performance in certain weather situations or for certain periods of time, it is important to be aware of Simpson's paradox (Simpson, 1951). It describes the phenomenon that a certain effect appearing in several subsamples may not be found in a combination of these samples, or that the larger sample may even show the complete opposite effect.

For example, a forecast model can have superior skill over all four seasons, compared with another model, but still be worse when assessed over the whole year. Hamill and Juras (2006) showed this to be true for a synthetic data set of temperature forecasts on two islands. In this case, the climatologies of the two islands were so different that the values of performance measures were misleadingly improved. Fricker, Ferro, and Stephenson (2013) found that this spurious skill does not affect proper scores derived from scoring rules, but care should be taken when using scores derived from a contingency table that are not proper, and skill scores in general.

In general, it is recommended to use statistical significance testing in order to evaluate potential model improvements. Differences in scores are often very small and it is hard to judge if they are caused by genuine improvement or chaotic error growth. Geer (2016) investigate a version of the Student's t-test modified for multiple models and taking account of autocorrelation in the scores. They also found that in order to detect an improvement of 0.5%, at least 400 forecast fields on a global grid would be required. This confirms our findings from Section 6.3.2 that it is essential to carefully consider the experiment sample size in order to generate meaningful and robust results.

6.5 SUMMARY

In this chapter, a variety of methods to assess different aspects of forecast goodness were presented and discussed. Calibration errors can be diagnosed with the help of histograms, in both univariate and multivariate settings. It is recommended to use multiple such diagnostics, especially in the multivariate case, as different tools highlight different types of miscalibration.

Scoring rules provide information about the accuracy of a forecast and are valuable tools for comparing forecasting methods. In this context, only proper scores should be used, as they ensure that the forecast based on the best knowledge will receive the best score. There are many such scores available, with the CRPS and the ignorance score being among the most popular. However, only looking at the mean of one such score can be misleading, even if the underlying sample seems to be of sufficient size. Therefore, it is crucial to also provide information about the error of a mean score, and to base decisions about model preference on the evaluation of multiple scoring rules, if possible. If we do not want to compare models, but rather understand the behavior of a model, it can be helpful to use measures that are not necessarily proper. Especially skill scores and the ACC are widely used.

By adding appropriate weight functions to the CRPS and the ignorance score, it is possible to evaluate extreme event forecasts in a proper way. These weight functions can be designed to emphasize, for example, different parts of the climatological distribution. Scores for multivariate quantities not only give information about the calibration and sharpness of the forecast, but also assess the correct representation of the covariance structure between locations, forecast times, or variables. However, some of them have limitations and do not work well if the number of dimensions is large.

Given the multitude of available evaluation tools and scores, which are constantly growing due to new research and applications, it is essential to be aware of their properties and how to choose a suitable measure. To make sure that all aspects of a forecast's performance are addressed, a number of scores should be calculated and a quantification of the associated uncertainty given.

REFERENCES

Anderson, J. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, *9*, 1518–1530.

Bentzien, S., & Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1924–1934.

Bigelow, F. (1905). Application of mathematics in meteorology. *Monthly Weather Review*, *33*, 90–90.

Brent, R. (1973). *Algorithms for Minimization Without Derivatives*. Englewood Cliffs: Prentice-Hall.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Bröcker, J. (2008). Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review*, *136*, 4488–4502.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, *135*, 1512–1519.

Dawid, A. (1984). Statistical theory: the prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Ser. A*, *147*, 278–292.

Dawid, A., & Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, *27*, 65–81.

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., & Stull, R. B. (2006). Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research: Atmospheres*, *111*, D24307.

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

Diks, C., Panchenko, V., & Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, *163*, 215–230.

Elmore, K. (2005). Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Weather and Forecasting*, *20*, 789–795.

Feldmann, K., Scheuerer, M., & Thorarinsdottir, T. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review*, *143*, 955–971.

Ferro, C., Richardson, D., & Weigel, A. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, *15*, 19–24.

Fricker, T., Ferro, C., & Stephenson, D. (2013). Three recommendations for evaluating climate predictions. *Meteorological Applications*, *20*, 246–255.

Friederichs, P., & Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, *135*, 2365–2378.

Friederichs, P., & Thorarinsdottir, T. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, *23*, 579–594.

Geer, A. J. (2016). Significance of changes in medium-range forecast scores. *Tellus Ser. A*, *68*, 30229.

Gilleland, E. (2017). Spatialvx: spatial forecast verification. *R package version 6-1*.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*, 746–762.

Gneiting, T., Balabdaoui, F., & Raftery, A. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Ser. B*, *69*, 243–268.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

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

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

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*, 411–422.

Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, *7*, 1747–1782.

Gneiting, T., Stanberry, L., Grimit, E., Held, L., & Johnson, N. (2008). Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder). *Test*, *17*, 211–264.

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society Ser. B*, *14*, 107–114.

Grimit, E., Gneiting, T., Berrocal, V., & Johnson, N. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, *132*, 2925–2942.

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

Hamill, T. M., & Juras, J. (2006). Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, *132*, 2905–2923.

Hering, A., & Genton, M. (2011). Comparing spatial predictions. *Technometrics*, *53*, 414–425.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*, 559–570.

Holzmann, H., & Klar, B. (2017). Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics*, *11*, 2404–2431.

Jolliffe, I., & Primo, C. (2008). Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, *136*, 2133–2139.

Jolliffe, I., & Stephenson, D. (Eds.), (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Chichester, UK: John Wiley & Sons.

Jordan, A., Krüger, F., & Lerch, S. (2017). *Evaluating probabilistic forecasts with the R package scoring Rules*. https://arxiv.org/abs/1709.04743 (Accessed 26 January 2018).

Krüger, F., Lerch, S., Thorarinsdottir, T. L., & Gneiting, T. (2016). *Probabilistic forecasting and comparative model assessment based on Markov Chain Monte Carlo output*. https://arxiv.org/pdf/1608.06802.pdf (Accessed 26 January 2018).

Kruskal, J. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, *7*, 48–50.

Lahiri, S. (2003). *Resampling Methods for Dependent Data*. New York: Springer.

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences Discussions*, *11*, 1267–1277.

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

Liu, R. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, *18*, 405–414.

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, *104*, 718–734.

Matheson, J., & Winkler, R. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*, 1087–1096.

Mirzargar, M., & Anderson, J. (2017). On evaluation of ensemble forecast calibration using the concept of data depth. *Monthly Weather Review*, *145*, 1679–1690.

Miyakoda, K., Hembree, G., Strickler, R., & Shulman, I. (1972). Cumulative results of extended forecast experiments I. Model performance for winter cases. *Monthly Weather Review*, *100*, 836–855.

Murphy, A. (1970). The ranked probability score and the probability score: a comparison. *Monthly Weather Review*, *98*, 917–924.

Murphy, A. (1973a). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, *12*, 215–223.

Murphy, A. (1973b). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595–600.

Murphy, A. (1974). A sample skill score for probability forecasts. *Monthly Weather Review*, *102*, 48–55.

Murphy, A. (1992). Climatology, persistence, and their linear combination as standards of reference in skill scores. *Weather and Forecasting*, *7*, 692–698.

Murphy, A. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*, 281–293.

Murphy, A., & Epstein, E. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, *117*, 572–582.

Oksanen, J., Blanchet, F., Friendly, M., Kindt, R., Legendre, P., & McGlinn, D. (2017). *Vegan: community ecology package*.

Palmer, T. (2012). Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *138*, 841–861.

Perkins, S., Pitman, A., Holbrook, N., & McAneney, J. (2007). Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *Journal of Climate*, *20*, 4356–4376.

Pinson, P. (2013). Wind energy: forecasting challenges for its operational management. *Statistical Science*, *28*, 564–585.

Pinson, P., & Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, *96*, 12–20.

Core Team, R. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roulston, M., & Smith, L. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, *130*, 1653–1660.

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

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

Siegert, S. (2017). Simplifying and generalising Murphy's Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, *143*, 1178–1183.

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Ser B*, *13*, 238–241.

Smith, L., & Hansen, J. (2004). Extending the limits of ensemble forecast verification with the minimum spanning tree. *Monthly Weather Review*, *132*, 1522–1528.

Stephenson, D., Coelho, C. A. S., & Jolliffe, I. (2008). Two extra components in the Brier score decomposition. *Weather and Forecasting*, *23*, 752–757.

Strähl, C., & Ziegel, J. (2017). Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, *11*, 608–639.

Thorarinsdottir, T., Gneiting, T., & Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, *1*, 522–534.

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

Tsyplakov, A. (2013). *Evaluation of probabilistic forecasts: Proper scoring rules and moments*. http://ssrn.com/abstract=2236605 (Accessed 26 January 2018).

Wei, W., Balabdaoui, F., & Held, L. (2017). Calibration tests for multivariate Gaussian forecasts. *Journal of Multivariate Analysis*, *154*, 216–233.

Weijs, S., & van Nooijen, R.van de Giesen, N. (2010). Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Review*, *138*, 3387–3399.

Wilks, D. (2004). The minimum spanning tree histogram as verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, *132*, 1329–1340.

Wilks, D. (2011). Statistical Methods in the Atmospheric Sciences. Oxford: Elsevier Academic Press.

Wilks, D. (2017). On assessing calibration of multivariate ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *143*, 164–172.

Ziegel, J., & Gneiting, T. (2014). Copula calibration. *Electronic Journal of Statistics*, *8*, 2619–2638.