Received: 28 March 2019 | Revised: 23 August 2019 | Accepted: 16 September 2019 | Published on: 25 November 2019

DOI: 10.1002/qj.3667

RESEARCH ARTICLE

Quarterly Journal of the Royal Meteorological Society RMetS

Probabilistic temperature forecasting with a heteroscedastic autoregressive ensemble postprocessing model

Annette Möller[1] | Jürgen Groß[2]

[1]Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, Clausthal-Zellerfeld, Germany

[2]Institute for Mathematics and Applied Informatics, University of Hildesheim, Hildesheim, Germany

Correspondence
Annette Möller, Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, Clausthal-Zellerfeld, Germany.
Email: annette.moeller@tu-clausthal.de

Funding information
Deutsche Forschungsgemeinschaft, MO 3394/1-1; Hungarian National Research, Development and Innovation Office, NN125679

Abstract

Weather prediction today is performed with numerical weather prediction (NWP) models. These are deterministic simulation models describing the dynamics of the atmosphere, and evolving the current conditions forward in time to obtain a prediction for future atmospheric states. To account for uncertainty in NWP models it has become common practice to employ ensembles of NWP forecasts. However, NWP ensembles often exhibit forecast biases and dispersion errors, thus require statistical postprocessing to improve reliability of the ensemble forecasts. This work proposes an extension of a recently developed postprocessing model utilizing autoregressive information present in the forecast error of the raw ensemble members. The original approach is modified to let the variance parameter depend on the ensemble spread, yielding a two-fold heteroscedastic model. Furthermore, an additional high-resolution forecast is included into the postprocessing model, yielding improved predictive performance. Finally, it is outlined how the autoregressive model can be utilized to postprocess ensemble forecasts with higher forecast horizons, without the necessity of making fundamental changes to the original model. We accompany the new methodology by an implementation within the `R` package `ensAR` to make our method available for other researchers working in this area. To illustrate the performance of the heteroscedastic extension of the autoregressive model, and its use for higher forecast horizons we present a case-study for a dataset containing 12 years of temperature forecasts and observations over Germany. The case-study indicates that the autoregressive model yields particularly strong improvements for forecast horizons beyond 24 h.

KEYWORDS

autoregressive process, ensemble postprocessing, heteroscedastic model, high-resolution forecast, predictive probability distribution, spread-adjusted linear pool, spread-error correlation

1 | INTRODUCTION

Today, weather prediction is based on numerical weather prediction (NWP) models. Such models are deterministic in nature and represent the dynamical physics of the atmosphere by a set of differential equations. The current state of the atmosphere is evolved forward in time to predict future atmospheric states. The solutions strongly depend on the initial conditions and model formulations. Thus, NWP models suffer from several sources of uncertainties. Common practice

[Correction added on 20 October 2020, after first online publication: Projekt Deal funding statement has been added.]

in addressing these uncertainties is the use of ensemble prediction systems (EPS). The NWP model is run multiple times, each time with variations in the model parametrizations and/or initial and boundary conditions (Gneiting and Raftery, 2005; Leutbecher and Palmer, 2008).

A forecast ensemble can be viewed as a probabilistic forecast that allows us to assess forecast uncertainty (Palmer, 2002). However, in practice, NWP ensembles exhibit forecast biases and dispersion errors, and require statistical postprocessing to improve calibration and forecast skill by utilizing recently observed forecast errors and observations. An additional benefit is that statistical postprocessing can yield full predictive probability distributions (Gneiting and Raftery, 2005; Wilks and Hamill, 2007; Gneiting and Katzfuss, 2014).

Statistical postprocessing models have enjoyed increasing popularity and success during the last decades and a variety of models tailored to specific problems have been developed. Many of the recently proposed models are extensions and modifications of two generic state-of-the-art models, namely the Ensemble Model Output Statistics approach (EMOS: Gneiting *et al.*, 2005) and the Bayesian Model Averaging (BMA: Raftery *et al.*, 2005).

The original EMOS and BMA models were designed for Gaussian distributed weather quantities, and a variety of modifications for other weather quantities have been developed (see, e.g., Schefzik *et al.*, 2013; Gneiting and Katzfuss, 2014; Hemri *et al.*, 2014, for an overview).

Further, postprocessing models allowing us to incorporate inter-variable, spatial, or temporal dependence structures have gained increased interest (e.g. Berrocal *et al.*, 2007; Kleiber *et al.*, 2011; Pinson, 2012; Schuhen *et al.*, 2012; Möller *et al.*, 2013; Schefzik *et al.*, 2013; Baran and Möller, 2015; Feldmann *et al.*, 2015; Hemri *et al.*, 2015; Vrac and Friederichs, 2015; Wilks, 2015; Ben Bouallègue *et al.*, 2016; Möller *et al.*, 2016; Schefzik and Möller, 2018).

A general overview of various aspects of ensemble postprocessing can be found in Vannitsem *et al.* (2018).

In line with the need for models incorporating dependencies explicitly, Möller and Groß (2016) introduced a postprocessing model for Gaussian distributed weather quantities (such as temperature) that accounts for dependencies of the individual ensemble forecasts across time. In this regard, the model utilizes the autoregressive information present in the forecast error of the individual raw ensemble forecasts to set up corrected ensemble forecasts as well as a predictive distribution.

The work presented here extends the AR-EMOS model of Möller and Groß (2016) to be of heteroscedastic (or “non-homogeneous”) nature, meaning that the variance parameter of the model varies with the (empirical) ensemble spread. A postprocessing model using a heteroscedastic variance parameter accounts for the well-known spread-error correlation (Barker, 1991; Whitaker and Loughe, 1998) of forecast ensembles, stating that there is a positive association between the forecast error (or predictive skill) and the spread of the ensemble. The extended AR-EMOS model incorporates heteroscedasticity in “two directions,” namely across time (longitudinal) for each individual member, and across the ensemble members (cross-sectional), to account for the above-mentioned spread-error correlation. Therefore, the approach allows for features not possible with standard postprocessing models, such as fitting a predictive distribution based only on a single ensemble member. This feature is investigated on the basis of an additional high-resolution forecast added to the ensemble, which is known to improve predictive performance to a great extent. While in the original article of Möller and Groß (2016) the AR-EMOS model was only applied to 24 h ahead ensemble forecasts, in general it is also applicable to (arbitrary) other forecast horizons. In this follow-up work we explain how the AR-EMOS model can be used for other than 24 h ahead forecast horizons and present results on predictive performance.

The development of postprocessing models accounting for specific problems is a quite active area of research; however, not all of the software carrying out model fitting for these recently proposed methods is publicly available. Prominent examples of postprocessing software implemented under the statistical software environment `R` (R Core Team, 2019) are the state-of-the-art EMOS and BMA models (Fraley *et al.*, 2018; Yuen *et al.*, 2018), a recently developed heteroscedastic logistic model for ensemble postprocessing (Messner *et al.*, 2013; 2014; 2017), implemented in the package `crch` (Messner *et al.*, 2016), and an implementation of verification metrics to assess probabilistic forecasts in the package `scoringRules` (Jordan *et al.*, 2017).

In line with the need for publicly available postprocessing software, this follow-up work to the methodology presented in Möller and Groß (2016) is accompanied by an implementation within an `R` package called `ensAR`, which can currently be installed from the Git repository hosting service GitHub by following the link given in the references section (Groß and Möller, 2019) and for example, making use of the package `devtools` (Wickham and Chang, 2018).

A case-study for temperature forecasts of the European Centre for Medium-range Weather Forecasts (ECMWF) (Buizza *et al.*, 2007) over Germany is carried out to illustrate the performance and properties of the proposed heteroscedastic autoregressive postprocessing model.

2 | METHODS

2.1 | Individual ensemble member postprocessing

Suppose that (ensemble) forecasts are initialized at a fixed time point and predict a weather quantity a fixed time step ahead (forecast horizon), which is (for now) not greater than

24 h. In the following, upper case (Latin) letters refer to the underlying random variable, and lower case (Latin) letters to the respective observed value. If t denotes the time point (day and hour) for which the forecast is valid, the data consists of forecasts $x_1(t), \dots, x_M(t)$, and a matching observation $y(t)$ of the underlying weather quantity $Y(t)$, for $t = 1, \dots, T$. For example, if a forecast is initialized at 1200 UTC and predicts 18 h ahead, the forecast is valid at 0600 UTC of the following day. Given the initialization time of the forecasts is fixed (which would usually be the case) the data (observation and ensemble members) is a collection of evenly (24 h) spaced time series referring to the respective validation time point. Let

$$Z_m(t) := Y(t) - x_m(t) \tag{1}$$

be the time series of forecast errors of the individual ensemble members $x_m(t)$.

Möller and Groß (2016) found that the observed individual error series $z_m(t)$ can exhibit substantial autoregressive behaviour. The authors propose to utilize this residual autoregressive information to obtain a corrected (AR-adjusted) forecast ensemble and to define a predictive distribution based on this AR-adjusted ensemble forecast.

In this regard, it is assumed that each $\{Z_m(t)\}$ follows an autoregressive process of order p_m, denoted by AR(p_m), that is,

$$Z_m(t) - \alpha_m = \sum_{j=1}^{p_m} \beta_{m,j}[Z_m(t-j) - \alpha_m] + \epsilon_m(t),$$

where $\{\epsilon_m(t)\}$ is white noise with expectation $\mathrm{E}(\epsilon_m(t)) = 0$ and variance $\mathrm{Var}(\epsilon_m(t)) = \sigma_m^2$. Then the random variable $Y(t)$ representing the weather quantity can be written as

$$Y(t) = \widetilde{x}_m(t) + \epsilon_m(t) \ ,$$

where

$$\widetilde{x}_m(t) = x_m(t) + \alpha_m + \sum_{j=1}^{p_m} \beta_{m,j}[y(t-j) - x_m(t-j) - \alpha_m]$$

can be viewed as a "corrected" forecast member for $Y(t)$ based on the original ensemble member $x_m(t), x_m(t-1), \dots, x_m(t-p_m)$ at past time points up to and including t, and the observation $y(t-1), \dots, y(t-p_m)$.

Performing the described procedure for each ensemble member $x_m(t)$ individually yields an "AR-adjusted" or "corrected" forecast ensemble $\widetilde{x}_1, \dots, \widetilde{x}_M$.

This approach of obtaining a corrected forecast ensemble rather than a predictive probability distribution/density has a connection to so-called "member-by-member-postprocessing" (MBMP: Van Schaeybroeck and Vannitsem, 2015; Schefzik, 2017). MBMP approaches have gained increased interest, as they retain the dependence structure inherent in the original raw forecast ensemble, while this implicit dependence information is often lost when performing (univariate) postprocessing.

2.2 | Forecast error variance

The variance of the autoregressive process $\{Z_m(t)\}$ in Equation 1, $\mathrm{Var}(Z_m(t)) =: \gamma_m^2(t)$, is given as

$$\gamma_m^2(t) = \frac{\sigma_m^2}{1 - \beta_{m,1}\rho_m(1) - \dots - \beta_{m,p_m}\rho_m(p_m)}, \tag{2}$$

where $\rho_m(k)$ is the autocorrelation function of the process $\{Z_m(t)\}$ at lag k, see for example Cryer and Chan (2008, equation 4.3.31). In R, the autocorrelation function of an autoregressive moving average (ARMA) process can be computed with the function `ARMAacf`.

2.3 | Different forecast horizons

If the (ensemble) forecasts are less than or equal to 24 h ahead, it is obvious how to obtain the corrected ensemble forecasts $\widetilde{x}_m$ based on the AR-fit to the error series $Z_m(t)$, as the values $x_m(t)$, $x_m(t-1)$, $\dots$, $x_m(t-p_m)$, as well as $y(t-1)$, $\dots$, $y(t-p_m)$ are readily available at time point t. So the required parameters of the AR process can be directly estimated from the (observed) training series $z(t-s), \dots, z(t-2), z(t-1)$ of length s.

However, if the considered forecast horizon is in the interval (24h, 48h] and t is any time point (day and hour) for which the forecast(s) are valid, then the observed error $z(t-1)$ is not available, as $y(t-1)$ has not yet been observed. Therefore, a pre-processing step is introduced before the AR-adjusted ensemble can be obtained. To compensate for the unavailable observed error $z(t-1)$, an AR process is fitted in advance to $z(t-s), \dots, z(t-3), z(t-2)$, and $z(t-1)$ is predicted from the respective AR model fit.

After predicting $z(t-1)$, the complete error series $z(t-s), \dots, z(t-2), z(t-1)$ is available again, and is then processed as described in section 2.1 by re-fitting the respective AR model to the full series, yielding the AR-adjusted ensemble valid at time point t.

For forecast horizons greater than 48 h, the observed errors at more time steps than $t-1$ are missing (e.g. for 72 h $z(t-1)$ and $z(t-2)$ are missing, and so on). The same procedure can then be applied to predict the missing errors from a fitted AR-model based on the past errors still available.

For prediction of future time-series values from a time-series model fit see also Shumway and Stoffer (2006, Sect. 3.5). In R the function `ar.predict` can be used for prediction from an autoregressive model.

2.4 | Heteroscedastic autoregressive predictive distribution

Möller and Groß (2016) assume the predictive distribution for $Y(t)$ to be Gaussian, that is

$$Y(t) \mid x_1(t), \dots, x_M(t) \sim \mathcal{N}(\mu(t), \sigma^2(t)), \tag{3}$$

where $\mu(t)$ may be a function of the ensemble members and $\sigma^2(t)$ may be a function of the ensemble variance.

To account for the well-known spread-error correlation in ensemble forecasts, this follow-up work proposes an improved model for the predictive variance $\sigma^2(t)$ in a similar fashion as the variance term is defined in the EMOS model (Gneiting *et al.*, 2005).

The predictive mean $\mu(t)$ is defined as the average over the AR-adjusted ensemble members (as in the original AR-EMOS model)

$$\mu(t) = \frac{1}{M}\sum_{m=1}^{M} \widetilde{x}_m(t) \ . \tag{4}$$

However, for the *standard deviation (SD)* $\sigma(t)$ of the predictive distribution an extended model is now suggested which combines longitudinal (time series) and cross-sectional (ensemble forecast) variation. The longitudinal part is defined as

$$\sigma_1(t) = \sqrt{\frac{1}{M}\sum_{m=1}^{M} \gamma_m^2(t)},$$

thus employing the average of the variances $\gamma_m^2(t)$ obtained from the autoregressive processes $Z_m(t)$ associated with the respective m ensemble members. The variances $\gamma_m^2(t)$ are solely based on information from past forecast errors (up to t), see Equation 2. The cross-sectional part is defined by

$$\sigma_2(t) = \sqrt{\widetilde{S}^2(t)} \ ,$$

where $\widetilde{S}^2(t)$ is the empirical variance of the AR-corrected ensemble members at the current time point t.

An intuitive and simple approach to incorporate both parts into a model for the predictive *SD* is to employ a convex combination of the form

$$\sigma(t) = w\sigma_1(t) + (1-w)\sigma_2(t), \tag{5}$$

where $w \in [0,1]$ is a weight obtained by minimizing the continuous ranked probability score (CRPS) over an (additional) training period. A solely longitudinal variance model can be obtained as a special case by setting $w = 1$, a solely cross-sectional variance model (in the fashion of standard EMOS) by setting $w = 0$. The simple model Equation 5 introduced above yields good results in terms of predictive performance as will be seen in the subsequent case-study. However, other more involved variance models might be considered in future research.

The EMOS model is also known by the term non-homogeneous regression, due to a variance model that is non-constant (non-homogeneous/heteroscedastic) with respect to the spread in the ensemble. Therefore, the modification of the AR-EMOS model with the variance model Equation 5 incorporating the ensemble spread is called heteroscedastic AR-EMOS – following the EMOS nomenclature. As the AR-EMOS method is based on a time-series model, and does not only consider the variation in the ensemble spread (cross-sectional part), but also the variation across time (longitudinal part), the more general (and in statistics more common) term heteroscedastic is used rather than the term non-homogeneous.

2.5 | Estimation

To fit the AR-EMOS model presented in section 2.4 (with the basic methodology in section 2.1), first the model parameters $\alpha_m, \beta_{m,1}, \ldots, \beta_{m,pm}$ for each member x_m are estimated by fitting an AR(p_m) process to the observed error series $\{z_m\}$ from a rolling training period by Yule–Walker estimation as carried out by the function `ar`; see also Shumway and Stoffer (2006, Sect. 3.6). The order p_m is automatically selected by a modified Akaike information criterion (AIC) proceeding as if the required variance estimate were obtained by maximum likelihood; see function ar (R Core Team, 2019).

This procedure is invoked with a default training length of 90 days, which has been found appropriate by Möller and Groß (2016).

To obtain the predictive mean $\mu(t)$ and *SD* $\sigma(t)$ a second rolling training period is required to estimate the weight parameter w used in the heteroscedastic variance model Equation 5 such that the average CRPS with respect to the predictive distribution $N(\mu(t),\sigma^2(t))$ is minimized for the training period. In the case-study following later, the default for the additional training window is set to 30 days length. In this setting, the 90 training days to estimate the parameters of the AR process directly precede the 30 additional training days to estimate the weights.

Therefore, to estimate *all* model parameters a training period of in total 120 days (with the default choices) is required, which is employed as a rolling window throughout the available dataset. This means the very first (out-of-sample in time) verification day is, for example, in the case of 24 h ahead forecasts, the day directly following the first 120 training days.

2.6 | Postprocessing a single forecast

When there exists a single distinguished forecast $x_*(t)$, for example, the high-resolution forecast $x_{\text{hres}}(t)$ described below, it is still possible to obtain a corresponding predictive distribution by the described AR-EMOS method.

The parameters $\mu_*(t)$ and $\sigma_*(t)$ are in principle estimated in the same way as those corresponding to the regular members. However, the mean in Equation 4 and the longitudinal part of the variance formula in Equation 5 reduce to a single summand, and the second part of the variance model in Equation 5 becomes zero. Nonetheless, the variance γ_*^2 of the error series corresponding to the individual member $x_*(t)$ can still be computed – on the basis of past values and the AR-fit. Thus, the original AR-EMOS approach and its refined version presented here both allow us to estimate the variance parameter and fit a predictive distribution based only on a single ensemble forecast.

Note that in case standard EMOS postprocessing approaches, as for example, Gneiting *et al.* (2005), define the variance parameter as a function of the ensemble sample variance computed with respect to at least two ensemble members at some time point t, the predictive distribution cannot be estimated based only on a single ensemble member without further ado.

2.7 | Postprocessing the raw ensemble mean

In order to make efficient use of the exchangeability of the raw ensemble, one might think of applying the autoregressive correction described above to the raw ensemble mean $\overline{x}$ only, thereby deducing the predictive distribution from the single forecast $x_* = \overline{x}$. Such an approach would yield a reduced number of parameters to estimate. This procedure, though perfectly applicable in accordance with our method described in sections 2.1 to 2.4, has already been found to be inferior with respect to evaluation measures for predictive performance in the earlier work by Möller and Groß (2016). Analysis of the data described in section 4 leads to the same conclusions, so that this approach is not further pursued here.

2.8 | High-resolution forecast

In this section we describe how the high-resolution forecast known to improve predictive performance (e.g. Kann *et al.*, 2009; Gneiting, 2014; Persson, 2015) can be included into the AR-EMOS postprocessing model. We call the AR-EMOS model including the additional forecast an extended model.

Let again $x_1(t), \ldots, x_M(t)$ denote the forecast ensemble. However, this time the M members comprise the regular (exchangeable) forecasts described in section 2.1 and the additional high-resolution forecast $x_{\mathrm{hres}}(t)$. That means the (total) number M of forecasts utilized is actually increased by one. As described in section 2.1 it is again assumed that the forecast errors $Z_m(t), m = 1, \ldots, M$ follow an AR(p_m) process, yielding the AR-corrected ensemble $\widetilde{x}_1(t), \ldots, \widetilde{x}_M(t), \widetilde{x}_{\mathrm{hres}}(t)$, which is the basis for estimating mean and variance of the predictive distribution.

As the high-resolution forecast has somewhat different properties than the regular ensemble members, an apparent approach may be to treat them as two different groups with respect to the parameters of the predictive distribution. This course of action is quite common in ensemble postprocessing models: ensemble members belonging to a certain group are considered exchangeable, and thus can be assumed to share the same coefficients in the model (Gneiting, 2014).

To account for the above-mentioned groups, each parameter in the AR-EMOS model is defined as the (equally weighted) sum of the respective group-wise parameters, that is,

$$\mu(t) = \frac{1}{2}(\mu_{\mathrm{ens}}(t) + \mu_{\mathrm{hres}}(t)),$$
$$\sigma(t) = \frac{1}{2}(\sigma_{\mathrm{ens}}(t) + \sigma_{\mathrm{hres}}(t)).$$

Here, $\mu_{\mathrm{ens}}(t)$ is estimated as already stated in Equation 4. The parameters $\mu_{\mathrm{hres}}(t)$ and $\sigma_{\mathrm{hres}}(t)$ corresponding to the high-resolution forecast are estimated as described in the previous subsection.

Assigning each of the group-specific parameters fixed and equal weights is a first relatively straightforward approach for demonstrating the general idea. In the case-study it will be shown that this simple version already yields good results for predictive performance.

Of course this rather simple method can be modified to be more data-driven, that is using weights for the group-specific parameters directly estimated from data, for example, by minimum CRPS estimation. Furthermore, the approach based on two groups (regular ensemble members and high-resolution forecast) can be generalized to include multiple groups of exchangeable forecast members, which need not necessarily contribute equally to predictive performance. Such a more general setting would make it reasonable to estimate the weights for the group-specific parameters from data.

One possibility for a more general definition of an AR-EMOS group model can be accomplished for example by combining (group-wise) predictive distributions with the spread-adjusted linear pool already employed by Möller and Groß (2016) and shortly described in the following subsection.

2.9 | Combination of predictive distributions

Möller and Groß (2016) proposed to combine the predictive distribution of classical EMOS and AR-EMOS in a spread-adjusted linear pool (SLP: Gneiting and Ranjan, 2013). For the special case of combining $n = 2$ predictive distributions, the SLP combination has cumulative distribution function (CDF)

$$F(x) = w_1 G_1(x) + w_2 G_2(x), \quad G_l(x) = \Phi\left(\frac{x - \mu_l}{\sigma_l c}\right),$$

$l = 1{,}2$, where w_1 is a non-negative weight parameter, $w_2 = 1 - w_1$, and c is a strictly positive spread adjustment parameter. Here, Φ denotes the cumulative distribution function (CDF) of the standard normal distribution. The two distributions G_1 and G_2 can be fitted separately by postprocessing models of choice, and the weights are obtained by minimizing a verification score (specifically the CRPS, see Equation 6, section 3) over a training period, for fixed and given G_1, G_2.

The original approach was proposed with the aim to improve predictive performance by combining two predictive distributions coming from different sources (see also Gneiting and Ranjan, 2013; Baran and Lerch, 2015; 2016; 2018). In principle, these approaches may be extended to a (finite) number $n > 2$ of predictive distributions.

3 | TOOLS TO ASSESS PREDICTIVE PERFORMANCE

3.1 | Scoring rules

A common tool to assess the quality of probabilistic forecasts are (proper) scoring rules. They assign a scalar to a pair (y, F), where y is the verifying observation and F the forecasting distribution (Gneiting *et al.*, 2007; Gneiting and Raftery, 2007; Gneiting, 2011). Scoring rules are negatively orientated such that smaller values indicate better performance.

A well-known and popular score is the continuous ranked probability score (CRPS), assessing calibration and sharpness simultaneously. For a predictive distribution represented by its cumulative distribution function (CDF) $F(y)$ and observation y_{obs} the CRPS is given as

$$\text{CRPS}(F, y_{\text{obs}}) = \int_{-\infty}^{\infty} \{F(y) - \mathbf{1}(y \geq y_{\text{obs}})\}^2 dy, \quad (6)$$

where $\mathbf{1}(y \geq y_{\text{obs}})$ equals 1 if $y \geq y_{\text{obs}}$ and 0 otherwise; see also Wilks (2011, Sect. 8.5.1).

To have a comparison of predictive performance with respect to scoring rules addressing different aspects of the predictive distribution, we employ two further frequently used scoring rules in the case-study.

The Dawid–Sebastiani score (DSS) is based only on the first two moments of the predictive distribution F. If μ_F and σ_F^2 denote the mean and variance of F, the Dawid and Sebastiani (1999) score is given by

$$\text{DSS}(F, y_{\text{obs}}) = \frac{(y_{\text{obs}} - \mu_F)^2}{\sigma_F^2} + 2\ln(\sigma_F); \quad (7)$$

see also Gneiting and Katzfuss (2014).

The logarithmic or ignorance score is a local scoring rule evaluating the negative logarithm of the predictive probability density function (PDF) $f(y)$ (locally) at the verifying observation. The score is defined as

$$\text{IGN}(f, y_{\text{obs}}) = -\ln(f(y_{\text{obs}})); \quad (8)$$

see also Wilks (2011), and Gneiting and Katzfuss (2014).

In case the predictive density $f(y)$ is Gaussian with parameters μ and σ^2, the ignorance score and the DSS have an explicit linear relationship given by

$$\text{DSS} = 2 \cdot \text{IGN} - \ln(2\pi);$$

see, for example, Wilks (2011).

Because of the above-mentioned linear relationship between IGN and DSS in the Gaussian case, considering the ignorance score only yields additional information in a comparative study to non-Gaussian predictive distributions. Therefore, we compute the IGN score in addition to the DSS when we compare EMOS and AR-EMOS with the SLP combination (which is a mixture of Gaussians).

3.2 | Visual assessment

To visually assess calibration of a probabilistic forecast, verification rank histograms and PIT histograms are employed (Wilks, 2011).

Here, the verification rank histogram (VRH) or Talagrand diagram is used to assess a forecast ensemble $x_1, \ldots ,x_M$. It can be obtained by computing the rank of the observation y within the ensemble (for each forecast case). If the ensemble members $x_1, \ldots ,x_M$ and the observation y are statistically indistinguishable (exchangeable), the rank of the observation with respect to the ensemble members has a discrete uniform distribution on $\{1, \ldots , M + 1\}$. The VRH then plots the empirical frequency of the observation ranks.

To assess calibration of a full predictive probability distribution, the frequencies of the Probability Integral Transform (PIT) values are plotted in equidistant bins. An observation y can be interpreted as a random sample from the "true" distribution F for the respective weather quantity. If the predicted distribution F_0 is identical to F, then $p = F_0(y)$ can be considered as realization of a uniformly distributed random variable on [0,1] and the plot of the frequencies of the PIT values p results in a uniform histogram.

3.3 | Further verification measures

The variance of the PIT values provides further information on the dispersion properties of the predictive distribution, a neutral dispersion being indicated by a variance equal to $1/12 = 0.0833$, the variance of the uniform distribution on [0,1], see Gneiting and Ranjan (2013).

The root mean variance (RMV) is used as a sharpness measure of predictive probability distributions. A main principle of probabilistic forecasting is "maximizing the sharpness of the predictive distribution subject to calibration" (Gneiting and Katzfuss) (e.g. Gneiting and Katzfuss, 2014), therefore the sharpness should be investigated in conjunction with the calibration.

3.4 | Testing for improvement in predictive performance

The statistical relevance of improvement in the verification scores may be investigated by testing for equal predictive performance of the two considered methods with the Diebold–Mariano test for time series; see Gneiting and Katzfuss (2014).

Let $s_1(t)$, $s_2(t)$ denote the time series of score values (e.g. the CRPS, the DSS, or another verification score) obtained from two competing methods as for example EMOS and AR-EMOS, for a verification period of length T (say). Then the large-sample standard normal test statistic adapted from Diebold and Mariano (1995) is given as

$$S = \sqrt{T}\frac{\overline{d}}{\sqrt{\sum_{\tau=-(h-1)}^{h-1} \hat{\rho}_d(\tau)}} ,$$

where

$$\overline{d} = \frac{1}{T}\sum_{t=1}^{T} d(t), \quad d(t) = s_1(t) - s_2(t),$$

is the average score differential and

$$\hat{\rho}_d(\tau) = \frac{1}{T}\sum_{t=|\tau|+1}^{T} (d(t) - \overline{d})(d(t-|\tau|) - \overline{d})$$

are the empirical autocovariances. The index h refers to the truncation choice for the lags at which autocovariances are incorporated in case of h-step ahead forecasts. As h-step ahead forecast errors in theory exhibit at most dependencies at $h-1$ lags (Diebold and Mariano, 1995), a typical recommendation is $h = 1$ in the case of forecast horizons in the interval [1,24] hours, $h = 2$ in case of forecast horizons in the interval (24,48], and so on.

4 | APPLICATION TO ECMWF TEMPERATURE FORECASTS

4.1 | Data description and data pre-processing

The data considered for our case-study contains the 50 member forecast ensemble by the ECMWF, see e.g. Molteni *et al.*, 1996; Buizza *et al.*, 2007. The data consist of 24, 48 and 72 h ahead forecasts initialized at 1200 UTC for 2 m surface temperature in Germany along with the verifying observations at 187 different stations in the time period ranging from 1 January 2002 to 20 March 2014; see also Hemri *et al.* (2014). In addition, there is one high-resolution forecast and one control forecast.

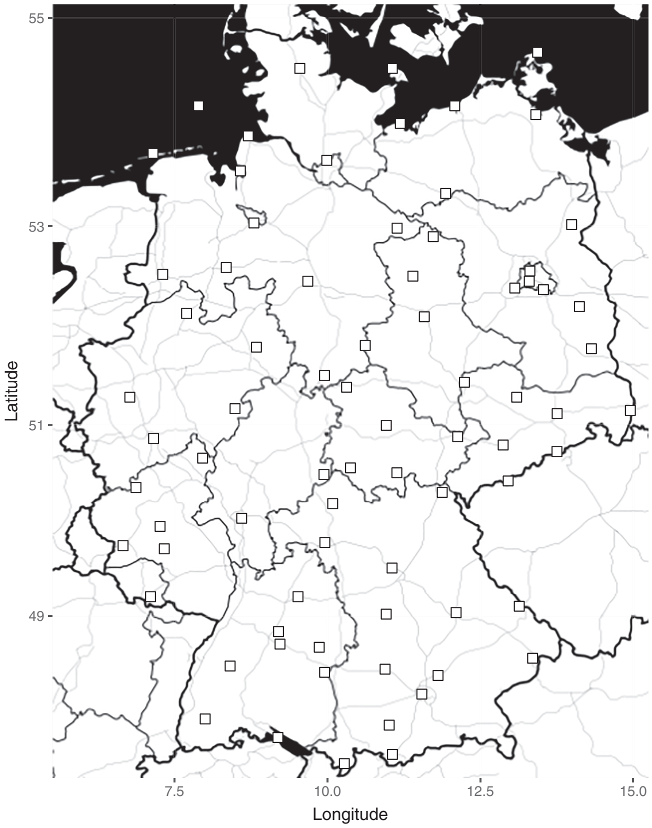

FIGURE 1 Seventy-six stations in Germany, chosen for admitting a modest occurrence of missing values in the raw dataset

For the application of time-series methods it is of importance to investigate whether the dates appear in chronological order and if some dates are missing. In the statistical software package `R`, missing values/cases are denoted by “NA” (not available).

From the full dataset with 187 stations, only those stations are retained which do not reveal NA gaps longer than 1. There are 76 stations which do match this rather strict specification. The new dataset still contains missing values.

The remaining missing values can in `R` be replaced by values obtained from linear interpolation using the function `na.approx` from the package `zoo` (Zeileis and Grothendieck, 2005).

Figure 1 shows the 76 stations retained for the subsequent analysis, where the station map was produced with the `R` package `ggmap` (Kahle and Wickham, 2013). In the subsequent analysis the station Magdeburg in eastern Germany will be considered for illustration purposes in the case-study.

To find possible outliers in the data, the test statistic from Chang *et al.* (1988) for detecting additive outliers is applied, being implemented as function `detectAO` in the `R` package `TSA` (Chan and Ripley, 2018). It requires the fit of an Autoregressive Integrated Moving Average (ARIMA) model to the series, which can, for example, be achieved by the function `auto.arima` from the `R` package `forecast` (Hyndman and Khandakar, 2008).

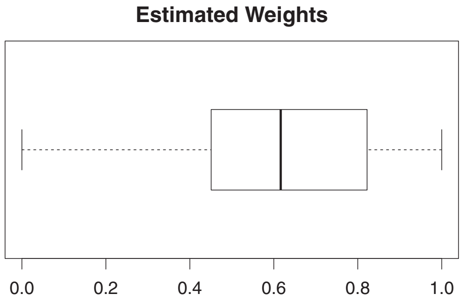

FIGURE 2 Estimated weights w associated with longitudinal part $\sigma_1(t)$ of the predictive *SD* $\sigma(t)$ for station Magdeburg

In order to find strong outlying observations, the significance level in `detectAO` is put to the very small value $\alpha = 0.00001$. By applying the above procedure to each station, it is found that six stations reveal suspicious values. From these, only two observations are removed from the dataset. At date 20 January 2003 and station Hannover the temperature observation is −90.8 °C, which is clearly impossible. At date 23 November 2002 and station Nürnberg, the temperature observation is −20.1 °C, which is very unusual with respect to preceding and succeeding temperature values and, in addition, is by far the smallest value in the complete series.

Removal is done by setting the outlying value to NA and then applying linear interpolation.

4.2 | Comparison of EMOS and heteroscedastic AR-EMOS

First, the state-of-the-art EMOS model is compared to the heteroscedastic AR-EMOS model (called AR-EMOS in the following) presented in section 2.4, where each model is based on the 50 regular ECMWF 24 h ahead ensemble forecasts. The parameters of the postprocessing models are estimated station-wise, based only on the data available for the respective station (so-called local approach).

A first analysis investigates the performance of the different models at the station Magdeburg in eastern Germany. Then, in a second step the analysis is carried out for all 76 stations in the dataset.

The AR-EMOS model is fitted as described in section 2.5. The EMOS model is fitted with the `R` package `ensembleMOS`. Estimating parameters of EMOS usually requires a training period of length between 20 to 40 days (Gneiting, 2014, Sect. 4), where in the subsequent study 30 days are chosen.

4.2.1 | Comparison at a single station

Figure 2 shows a boxplot of the weights w associated with the longitudinal part $\sigma_1(t)$ of $\sigma(t)$ from Equation 5. The median is at 0.617.

The plots displayed in Figure 3 show a comparison of the predictive mean and *SD* of the EMOS and (heteroscedastic) AR-EMOS predictive distribution for Magdeburg. The predictive means of both methods exhibit a strong relationship (squared correlation coefficient equal to 0.9963), while the predictive *SD* differ to a certain extent (squared correlation coefficient equal to 0.5142). So, although AR-EMOS proceeds in a quite different way to estimate the predictive mean, the result does apparently not differ from EMOS very much. The different approaches to estimating the variance obviously also yield different results; the AR-EMOS *SD* have a tendency to be smaller than those of EMOS. The boxplots for the *SD* also show that for EMOS there is much more variation in the estimated *SD* than for AR-EMOS.

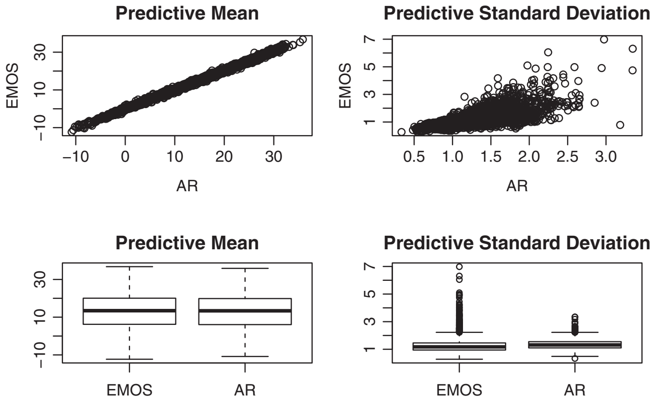

FIGURE 3 Comparison of predictive mean and predictive *SD* obtained by EMOS and (heteroscedastic) AR-EMOS for station Magdeburg

TABLE 1 Verification metrics of EMOS and (heteroscedastic) AR-EMOS for station Magdeburg aggregated over 4,341 verification dates

	CRPS	DSS	RMV	Var(PIT)
EMOS	0.8415	2.0918	1.3670	0.0946
AR-EMOS	0.8309	1.9149	1.3825	0.0876

Table 1 presents the CRPS, the DSS, the root mean variance (RMV) and the PIT variance for both methods at station Magdeburg. The top row in Figure 4 additionally shows the PIT histograms of both methods at station Magdeburg.

The PIT values of both models have a variance greater than 1/12 = 0.0833, indicating under-dispersion of the predictive distributions. This under-dispersion is visible in the respective PIT histograms as well. However, the PIT variance of AR-EMOS is much closer to 1/12 than the PIT variance of EMOS. When looking at the PIT histograms, the EMOS histogram indicates a slightly more pronounced bin for small PIT values, indicating a stronger forecast bias. On the contrary, the EMOS predictive distribution is slightly sharper than the AR-EMOS one, however obviously at the expense of dispersion accuracy.

When looking at the verification scores providing an overall judgment on predictive performance (CRPS, DSS), we can

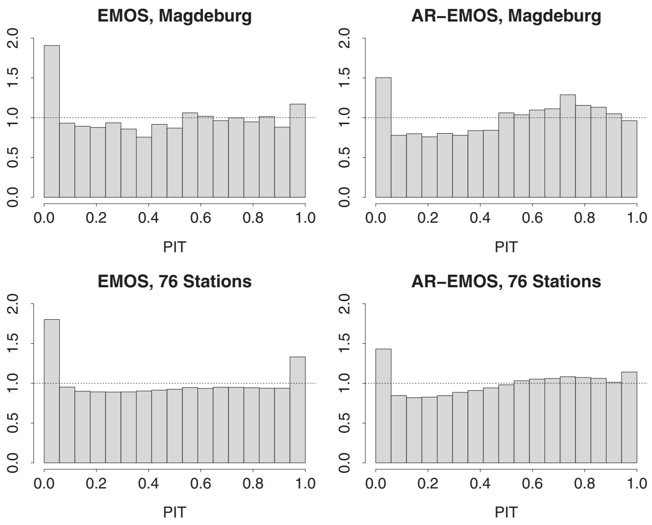

FIGURE 4 PIT histograms of EMOS and (heteroscedastic) AR-EMOS aggregated over 4,341 verification days (top row), as well as PIT histograms of both methods aggregated over 4,341 × 76 verification cases (dates and stations, bottom row)

further conclude that AR-EMOS performs better than EMOS with respect to the average CRPS as well as to the average DSS at station Magdeburg.

As the difference in CRPS values is relatively small, we investigate whether AR-EMOS provides a statistically significant improvement in CRPS over state-of-the-art EMOS, by a one-sided Diebold–Mariano test for the alternative

$$H_1 : CRPS_{\text{AR-EMOS}} < CRPS_{\text{EMOS}}.$$

For station Magdeburg, the resulting p-value is given as .01722. Thus, the test shows that the CRPS values of AR-EMOS are (on average) indeed significantly smaller than those of EMOS.

4.2.2 | Comparison at all stations

In a second step the analysis performed for a single example station is carried out for all 76 stations in the dataset, and the results are aggregated. For this, a local approach is used, that is, the considered models are estimated at each station individually, resulting in location-specific model parameters based only on the data of a specific station.

When aggregating over all 76 stations, for each station 4,341 verification days are available, yielding in total 329,916 forecast cases to aggregate over. Table 2 shows the resulting verification metrics for EMOS and AR-EMOS.

TABLE 2 Verification metrics of EMOS and (heteroscedastic) AR-EMOS aggregated over 4,341 × 76 verification cases (dates and stations)

	CRPS	DSS	RMV	Var(PIT)
EMOS	0.9057	2.1523	1.4907	0.0954
AR-EMOS	0.9033	2.0129	1.5322	0.0883

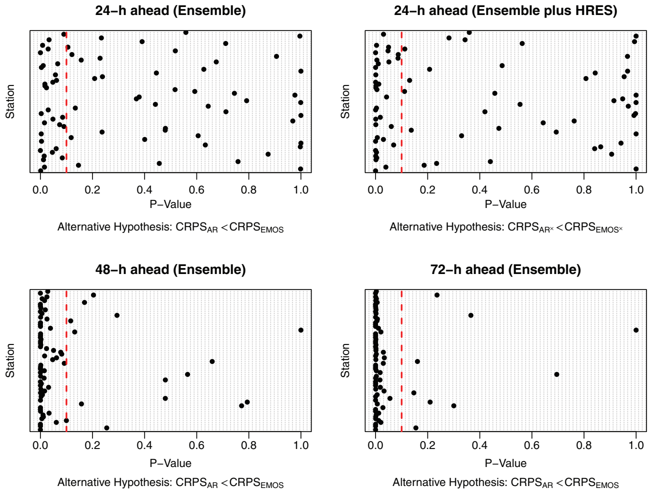

FIGURE 5 Station-wise p-values of the Diebold–Mariano one-sided test comparing EMOS and (heteroscedastic) AR-EMOS. (Truncation index $h = 1$ for 24 h ahead, $h = 2$ for 48 h ahead, and $h = 3$ for 72 h ahead forecasts)

The verification scores indicate that the predictive distribution of AR-EMOS has comparable but slightly different properties than EMOS, similar to the comparison at a single station. While CRPS and DSS of AR-EMOS are (slightly) smaller than those of EMOS, the RMV is slightly larger, indicating the AR-EMOS predictive distribution to be less sharp than its EMOS counterpart. However, the sharper distribution of EMOS comes at the expense of calibration, the PIT variance of EMOS is much larger than 1/12, indicating under-dispersion, while the PIT variance of AR-EMOS is relatively close to 1/12, indicating a distribution with dispersion properties close to neutral dispersion. These observations are consistent with the PIT histograms in (the bottom row of) Figure 4, where the EMOS PIT histogram clearly exhibits a U-shape, with indicates under-dispersion, while the U-shape is much less pronounced in the AR-EMOS PIT histogram.

To find statistical evidence about the significance of the difference in predictive performance between the two methods, the one-sided Diebold–Mariano test is conducted again, this time for the CRPS time series at each of the 76 stations individually.

The upper-left panel of Figure 5 shows the station-wise p-values of the Diebold–Mariano test for EMOS vs. AR-EMOS. Small p-values give statistical evidence for the alternative hypothesis that the values of the AR-EMOS CRPS series are on average smaller than the values of the EMOS CRPS series, thus indicating superior performance of AR-EMOS compared to EMOS. At 31 stations the p-value is $\leq .1$, thus indicating superior performance of AR-EMOS; see the vertical red dashed line in (the upper left panel of) Figure 5.

TABLE 3 Verification metrics of EMOS*, (heteroscedastic) AR-EMOS*, and SLP combination SLP* of both, aggregated over 4,251 verification dates at station Magdeburg

	CRPS	IGN	DSS	RMV	Var(PIT)
EMOS*	0.8223	1.9341	2.0304	1.3714	0.0908
AR-EMOS*	0.8097	1.8391	1.8404	1.3663	0.0849
SLP*	0.8000	1.8598	1.9043	1.3965	0.0854

4.3 | Incorporating the high-resolution forecast with group approach

The ECMWF ensemble also comprises a single high-resolution run, whose importance for statistical postprocessing is described, for example, by Gneiting (2014). As indicated in section 2.8, extended postprocessing models are now considered, which include the high-resolution forecast.

For the ECMWF data considered here, there are 50 exchangeable (that is, statistically indistinguishable) forecast members forming one group, while the high-resolution forecast is regarded as a second group due to its different properties.

Within the ensembleMOS package, the group membership of each ensemble forecast can be directly specified. As described in section 2.8, a straightforward ad hoc way to implement a 2-group AR-EMOS model is to represent the model parameters as a sum of the two group-specific parameters and assign the group-specific parameters (fixed) equal weight.

Furthermore, the SLP combination of EMOS and AR-EMOS proposed in Möller and Groß (2016) is revisited. However, in contrast to the original analysis, here the SLP combination of EMOS and AR-EMOS based on the 50 exchangeable members *and* the additional high-resolution forecast is investigated (which we called extended models, denoted by EMOS* and AR-EMOS*, respectively).

As additional training data are needed to estimate the weights in the SLP combination, the final number of verification days considered differs from the above analyses comparing only EMOS and AR-EMOS. Here, the results at the station Magdeburg are aggregated over 4,251 verification days. When aggregating over all 76 stations (each with 4,251 verification days) as well, the results are based on 323,076 forecast cases in total.

Results for verification scores at the station Magdeburg are presented in Table 3. It is clearly visible that in terms of CRPS, ignorance score and DSS AR-EMOS* improves over EMOS* to a large extent. The improvement in CRPS and DSS is much more pronounced than in the case where the high-resolution forecast was not incorporated into both models. The SLP* combination of the two models improves the CRPS even more in comparison to EMOS*.

Concerning sharpness as measured by the RMV, the AR-EMOS* model yields the sharpest predictive distribution, with a PIT variance extremely close to 1/12 at the same time. EMOS* and the SLP* combination of both models are less sharp (with EMOS* being slightly sharper than SLP*); however, while EMOS* has a PIT variance indicating under-dispersion (larger than 1/12), the SLP* combination has a PIT variance close to neutral dispersion. Therefore, at the station Magdeburg, the sharpness-calibration properties of AR-EMOS* seem to be appropriate and better than those of the other predictive distributions.

Although the improvement in CRPS of AR-EMOS* compared to EMOS* is much more obvious as in the respective analysis at Magdeburg presented in Table 1, a one-sided Diebold–Mariano test for $H_1 : CRPS_{\text{AR-EMOS}^*} < CRPS_{\text{EMOS}^*}$ at Magdeburg is performed to investigate the significance of the improvement. The resulting p-value is .01233, showing that AR-EMOS* is indeed performing significantly better than EMOS* at Magdeburg in terms of CRPS.

4.3.1 | Comparison of all stations

Next, the above described comparison of EMOS*, AR-EMOS* and SLP* is conducted for all 76 stations, where again the models are estimated station-wise. Due to the need for additional training data for the SLP combination, the number of verification cases considered differs from the analysis presented in Table 2 as alreadymentioned for the station Magdeburg. Here, the aggregation in Table 4 is performed over 323,076 verification cases (4,251 verification dates at each of the 76 stations). The aggregated verification metrics show that AR-EMOS* performs better than EMOS* with respect to the CRPS, ignorance score and DSS. With respect to the CRPS and ignorance score the SLP combination performs best; with respect to DSS it also performs clearly better than EMOS*. EMOS* has the sharpest predictive distribution in terms of the RMV, while AR-EMOS* and SLP* exhibit a similar level of sharpness. However, the PIT variance of AR-EMOS* is much closer to that of neutral dispersion than EMOS*, having a PIT variance indicating under-dispersion.

Station-wise p-values of the one-sided Diebold–Mariano test for the CRPS series are computed to investigate whether the improvement in CRPS is significant. The upper-right panel of Figure 5 shows the resulting p-values at all 76 stations for the one-sided Diebold–Mariano test for the alternative $H_1 : CRPS_{\text{AR-EMOS}^*} < CRPS_{\text{EMOS}^*}$.

TABLE 4 Verification metrics of EMOS*, (heteroscedastic) AR-EMOS*, and SLP combination SLP* of both, aggregated over 4,251 × 76 verification cases (dates and stations)

	CRPS	IGN	DSS	RMV	Var(PIT)
EMOS*	0.8712	1.9896	2.1412	1.4250	0.0931
AR-EMOS*	0.8685	1.8825	1.9270	1.4950	0.0854
SLP*	0.8460	1.8782	1.9350	1.5031	0.0860

TABLE 5 Verification metrics of EMOS(ENS) and (heteroscedastic) AR-EMOS(ENS) aggregated over 4,340 × 76 (48 h) and 4,339 × 76 (72 h) verification cases (dates and stations)

	CRPS	DSS	RMV	Var(PIT)
EMOS 48 h	1.0101	2.4147	1.6156	0.0979
AR-EMOS 48 h	0.9897	2.1749	1.7263	0.0872
EMOS 72 h	1.1244	2.6353	1.7831	0.099
AR-EMOS 72 h	1.0949	2.3548	1.9591	0.086

Again, the dashed red line denotes the significance level 0.1. When incorporating the high-resolution forecast into the models, the number of stations where AR-EMOS* performs significantly better (at significance level 0.1) than EMOS* increases to 41.

4.4 | Post-processing for higher forecast horizons

Finally we analyze the performance of the heteroscedastic AR-EMOS model for higher forecast horizons. In many applications, typically 24 h ahead forecasts are investigated, but higher forecast horizons are often not considered. To illustrate the effect, we present results for 48 and 72 h ahead forecasts.

In section 2.3 the procedure for applying AR-EMOS to forecast horizons greater than 24 h is explained. EMOS is capable of dealing with other than 24 h ahead forecasts as well, and the forecast horizon considered can be explicitly specified within the ensembleMOS package.

Table 5 shows the verification metrics for EMOS and AR-EMOS, based on 48 and 72 h ahead ensemble forecasts. For 48 h ahead forecasts the verification metrics and PIT histograms are based on a total of 329,840 verification cases (4,340 verification days for each of the 76 stations); for 72 h ahead forecasts, they are based on 329,764 verification cases (4,339 verification days for each station).

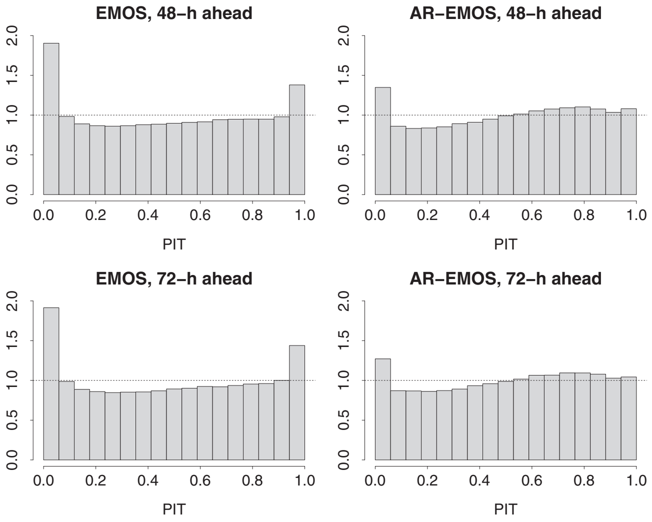

FIGURE 6 PIT histograms of EMOS and (heteroscedastic) AR-EMOS for 48 h ahead (top row), and for 72 h ahead (bottom row), aggregated over 76 stations, each with 4,340 (48 h) and 4,339 (72 h) verification dates

For both forecast horizons, it is clearly visible that AR-EMOS improves on EMOS in terms of CRPS and DSS, with the improvement being even more pronounced for 72 h ahead forecasts. Compared to the results on 24 h ahead forecasts, the improvement of AR-EMOS over EMOS becomes clearer the larger the forecast horizon. For both considered horizons in Table 5, the EMOS predictive distribution is a bit sharper than its AR-EMOS counterpart; however, in each case the PIT variance of EMOS indicates under-dispersion to a larger extent than the PIT variance of AR-EMOS.

Figure 6 presents the respective PIT histograms of EMOS and AR-EMOS, where the top panel refers to 48 h ahead, and the bottom panel to 72 h ahead forecasts.

To investigate whether the improvement of AR-EMOS over EMOS is indeed a significant one, again the Diebold–Mariano test is performed at each station for the same one-sided alternative as in the previous paragraphs. The lower panel of Figure 5 displays the resulting p-values for 48 h (left panel) and 72 h (right panel). For both forecast horizons the improvement in predictive performance of AR-EMOS over EMOS in terms of CRPS is highly significant for most of the stations: for 48 h ahead forecasts the p-value is $\leq .1$ at 62 stations (and $\leq .05$ still at 55 stations), for 72 h ahead forecasts even at 67 stations (and $\leq .05$ at 66 stations).

When moving to higher forecast horizons, the number of stations where AR-EMOS is significantly superior to EMOS increases heavily, and at more and more stations the level of significance even gets smaller. This indicates that the performance of AR-EMOS increases in comparison to EMOS for higher forecast horizons.

5 | CONCLUSION

This follow-up work presents some new features and extensions of the AR-EMOS model introduced by Möller and Groß (2016), and is accompanied by an implementation of the method within an R package called ensAR (Groß and Möller, 2019). The original model for the predictive variance is extended to incorporate the ensemble spread, yielding a heteroscedastic model implicitly accounting for the spread-error correlation, in a slightly different way than the EMOS model. The (heteroscedastic) AR-EMOS model allows us to fit a predictive distribution to a single ensemble member, as the longitudinal part of the (extended) predictive variance can still be computed for a single ensemble forecast, which is an advantage over standard postprocessing approaches such as EMOS.

Additionally, incorporation of a high-resolution forecast is investigated. The conducted case-study indicates that this forecast improves predictive performance to a large extent. To incorporate the high-resolution forecast, an AR-EMOS group model is defined, which follows a somewhat different approach than the EMOS group model. In the case-study, only a simple heuristic form of the AR-EMOS group model is considered, which already yields excellent results. However, extensions to a more general and data-driven form are relatively straightforward and subject to future research.

Finally, a feature of the AR-EMOS model not discussed in the original work is presented. The model allows us to fit predictive distributions based on ensemble forecasts with arbitrary forecast horizons. In the original work a case-study based only on 24 h ahead forecasts is presented. However, the AR-EMOS model can postprocess ensemble forecasts with arbitrary forecast horizons. For forecast horizons smaller or equal to 24 h ahead the model can be directly employed without any additional modifications. For horizons larger than 48 h ahead, the model can be applied by adding only one small pre-processing step, namely predicting the days between the last validation date of the forecast and the verification date with the AR model, also used to set up the AR-EMOS method itself.

The conducted case-study indicates that for forecast horizons beyond 24 h ahead (with 48 and 72 h ahead considered as examples) the AR-EMOS performs particularly well, and improves significantly over EMOS. Therefore, the autoregressive postprocessing approach shows potential for accurate prediction at higher forecast horizons.

While in this work and in the original article by Möller and Groß (2016) a univariate time series approach was employed, multivariate time-series models, such as vector autoregressive (VAR) processes, may be investigated in future research. Possible multivariate settings of interest could for example involve modelling the forecast errors of each ensemble member jointly, modelling dependencies between ensemble forecasts and observations, or modelling dependencies between observation locations by spatial time-series models, or in a more general setting, within the framework of space–time models.

Furthermore, the forecast error was investigated for autoregressive behaviour only. However, considering other (more general) stochastic processes might be beneficial in the application of ensemble postprocessing as well, and will be part of future research. For example, more general ARIMA models could be considered, or Generalized AutoRegressive Conditional Heteroscedasticity (GARCH) models to account explicitly for heteroscedastic variances in the ensemble forecasts. Furthermore, estimating Markov processes for the forecast errors might be beneficial. In line with the recently upcoming interest in application of machine learning approaches to ensemble postprocessing (e.g. Rasp and Lerch, 2018), application of more data-driven methods such as neural networks for time-series data could provide a highly data-adaptive alternative to standard time-series models.

ACKNOWLEDGEMENTS

The authors acknowledge support of the research by Deutsche Forschungsgemeinschaft (DFG) Grant No. MO 3394/1-1, and by the Hungarian National Research, Development and Innovation Office under Grant No. NN125679. Furthermore, we are grateful to the European Centre for Medium-Range Weather Forecasts (ECMWF) and the German Weather Service (DWD) for providing forecast and observation data, respectively.

The authors wish to thank two anonymous referees for valuable comments and suggestions, which helped to improve the presentation considerably.

Open access funding enabled and organized by Projekt DEAL.

ORCID

Annette Möller https://orcid.org/0000-0001-9386-1691

REFERENCES

Baran, S. and Lerch, S. (2015) Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141, 2289–2299.

Baran, S. and Lerch, S. (2016) Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, 27, 116–130.

Baran, S. and Lerch, S. (2018) Combining predictive distributions for statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34, 447–496.

Baran, S. and Möller, A. (2015) Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics*, 26, 120–132.

Barker, T. (1991) The relationship between spread and forecast error in extended-range forecasts. *Journal of Climate*, 4, 733–742.

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. *Monthly Weather Review*, 144, 4737–4750.

Berrocal, V., Raftery, A. and Gneiting, T. (2007) Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135, 1386–1402.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G. and Vitart, F. (2007) The new ECMWF VAREPS (variable resolution ensemble prediction system). *Quarterly Journal of the Royal Meteorological Society*, 133, 681–695.

Chan, K.-S. and Ripley, B. (2018) *TSA: time series analysis*. R package version 1.2. Available at: https://CRAN.R-project.org/package=TSA [Accessed 18th August 2015].

Chang, I., Tiao, G.C. and Chen, C. (1988) Estimation of time series parameters in the presence of outliers. *Technometrics*, 30, 193–204.

Cryer, J.D. and Chan, K.-S. (2008) *Time Series Analysis: With Applications in R*. New York, NY: Springer.

Dawid, A.P. and Sebastiani, P. (1999) Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 27, 65–81.

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

Feldmann, K., Scheuerer, M. and Thorarinsdottir, T. (2015) Spatial postprocessing of ensemble forecasts for temperature using non-homogeneous Gaussian regression. *Monthly Weather Review*, 143, 955–971.

Fraley, C., Raftery, A.E., Sloughter, J.M., Gneiting, T. and University of Washington. (2018) *ensembleBMA: probabilistic forecasting using ensembles and Bayesian model averaging*. R package version 5.1.5. Available at: http://CRAN.R-project.org/package=ensembleBMA [Accessed 18th August 2015].

Gneiting, T. (2011) Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762.

Gneiting, T. (2014) *Calibration of medium-range weather forecasts*. ECMWF Technical Memorandum 719. Reading, UK: European Centre for Medium-Range Weather Forecasts. Available at: http://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf.

Gneiting, T., Balabdaoui, F. and Raftery, A. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69, 243–268.

Gneiting, T. and Katzfuss, M. (2014) Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1, 125–151.

Gneiting, T. and Raftery, A. (2005) Weather forecasting with ensemble methods. *Science*, 310, 248–249.

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

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

Gneiting, T. and Ranjan, R. (2013) Combining predictive distributions. *Electronic Journal of Statistics*, 7, 1747–1782.

Groß, J. and Möller, A. (2019) *ensAR: autoregressive postprocessing methods for ensemble forecasts*. R package version 0.2.0. Available at: https://github.com/JuGross/ensAR [Accessed 18th August 2015].

Hemri, S., Lisniak, D. and Klein, B. (2015) Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, 51, 7436–7451.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. and Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41, 9197–9205. https://doi.org/10.1002/2014GL062472.

Hyndman, R. and Khandakar, Y. (2008) Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26, 1–22. Available at: http://ideas.repec.org/a/jss/jstsof/27i03.html [Accessed 4 March 2019].

Jordan, A., Krüger, F. and Lerch, S. (2017) Evaluating probabilistic forecasts with scoringRules. *arXiv preprint*, arXiv, 1709.04743.

Kahle, D. and Wickham, H. (2013) ggmap: spatial visualization with ggplot2. *The R Journal*, 5, 144–161. Available at: https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf [Accessed 4 March 2019].

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, 3373–3386.

Kleiber, W., Raftery, A., Baars, J., Gneiting, T., Mass, C. and Grimit, E. (2011) Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Monthly Weather Review*, 139, 2630–2649.

Leutbecher, M. and Palmer, T.N. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539.

Messner, J.W., Mayr, G.J. and Zeileis, A. (2016) Heteroscedastic censored and truncated regression with crch. *The R Journal*, 8, 173–181. https://doi.org/10.32614/RJ-2016-012.

Messner, J.W., Mayr, G.J. and Zeileis, A. (2017) Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145, 137–147.

Messner, J.W., Mayr, G.J., Zeileis, A. and Wilks, D.S. (2014) Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review*, 142, 448–456.

Messner, J.W., Zeileis, A., Broecker, J. and Mayr, G.J. (2013) Probabilistic wind power forecasts with an inverse power curve transformation and censored regression. *Wind Energy*, 17, 1753–1766.

Möller, A. and Groß, J. (2016) Probabilistic temperature forecasting based on an ensemble autoregressive modification. *Quarterly Journal of the Royal Meteorological Society*, 142(696), 1385–1394. https://doi.org/10.1002/qj.2741.

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.

Möller, A., Thorarinsdottir, T., Lenkoski, A. and Gneiting, T. (2016) Spatially adaptive, Bayesian estimation for probabilistic temperature forecasts. *arXiv preprint*, 15. Available at: http://arxiv.org/abs/arXiv:1507.05066.

Molteni, F., Buizza, R., Palmer, T.N. and Petroliagis, T. (1996) The ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122, 73–119.

Palmer, T.N. (2002) The economic value of ensemble forecasts as a tool for risk assessment: from days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128, 747–774.

Persson, A. (2015) *User guide to ECMWF products*. Available at: http://www.ecmwf.int/en/forecasts/documentation-and-support [Accessed 22 May 2018].

Pinson, P. (2012) Adaptive calibration of (u,v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138, 1273–1284.

R Core Team. (2019) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ [Accessed 18 August 2019].

Raftery, A., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174.

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

Schefzik, R. (2017) Ensemble calibration with preserved correlations: unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Quarterly Journal of the Royal Meteorological Society*, 143, 999–1008.

Schefzik, R. and Möller, A. (2018) Ensemble postprocessing methods incorporating dependence structures. In: Vannitsem, S., Wilks, D. and Messner, J. (Eds.) *In Statistical Postprocessing of Ensemble Forecasts*. Amsterdam: Elsevier, pp. 91–125.

Schefzik, R., Thorarinsdottir, T. and Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28, 616–640.

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

Shumway, R. and Stoffer, D. (2006) *Time Series Analysis and its Applications: With R examples*, 2nd edition. New York: Springer.

Van Schaeybroeck, B. and Vannitsem, S. (2015) Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141, 807–818.

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

Vrac, M. and Friederichs, P. (2015) Multivariate – inter-variable, spatial, and temporal – bias correction. *Journal of Climate*, 28, 218–237.

Whitaker, J. and Loughe, A. (1998) The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, 126, 3292–3302.

Wickham, H. and Chang, W. (2018) *devtools: tools to make developing R packages easier*. R package version 2.0.1. Available at: https://CRAN.R-project.org/package=devtools [Accessed 18 August 2019].

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*. Amsterdam: Academic Press.

Wilks, D. (2015) Multivariate ensemble model output statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, 141, 945–952.

Wilks, D. and Hamill, T. (2007) Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, 135, 2379–2390.

Yuen, R., Baran, S., Gneiting, T., Thorarinsdottir, T., Fraley, C., Lerch, S. and Scheuerer, M. (2018) *ensembleMOS: ensemble model output statistics*. R package version 0.8.2. Available at: http://CRAN.R-project.org/package=ensembleMOS [Accessed 18 August 2019].

Zeileis, A. and Grothendieck, G. (2005) zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14, 1–27. Available at: http://www.jstatsoft.org/v14/i06/ [Accessed 4 March 2019].

How to cite this article: Möller A, Groß J. Probabilistic temperature forecasting with a heteroscedastic autoregressive ensemble postprocessing model. *Q J R Meteorol Soc*. 2020;146:211–224. https://doi.org/10.1002/qj.3667