Simultaneous Ensemble Postprocessing for Multiple Lead Times with Standardized Anomalies

MARKUS DABERNIG AND GEORG J. MAYR
Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria

JAKOB W. MESSNER
Institute of Atmospheric and Cryospheric Sciences, and Department of Statistics, University of Innsbruck, Innsbruck, Austria

ACHIM ZEILEIS
Department of Statistics, University of Innsbruck, Innsbruck, Austria

(Manuscript received 28 October 2016, in final form 5 April 2017)

ABSTRACT

Separate statistical models are typically fit for each forecasting lead time to postprocess numerical weather prediction (NWP) ensemble forecasts. Using standardized anomalies of both NWP values and observations eliminates most of the lead-time-specific characteristics so that several lead times can be forecast simultaneously. Standardized anomalies are formed by subtracting a climatological mean and dividing by the climatological standard deviation. Simultaneously postprocessing forecasts between $1/12$ and $1/120$ h increases forecast coherence between lead times, yields a temporal resolution as high as the observation interval (e.g., up to 10 min), and speeds up computation times while achieving a forecast skill comparable to the conventional method.

1. Introduction

Weather forecasts are important for many aspects of life whether professional or private. These weather forecasts rely mainly on numerical weather prediction (NWP) models, which solve partial differential equations. However, it is impossible to solve these equations exactly and errors occur, which grow with forecasting horizon. Additionally, because of limited computational power, the numerical grid of these forecast models has to be coarse and subgrid processes have to be parameterized. The systematic component of the resulting errors can be corrected with a statistical model trained on numerical forecasts and their matching observations of previous forecast times (Glahn and Lowry 1972). Errors and resulting corrections are larger over complex terrain, such as the Alps, where real and NWP model topography differ strongly.

Additionally, weather centers compute not one but an ensemble of NWP forecasts in order to account for uncertainties introduced in initial conditions and parameterizations (Lorenz 1982; Leutbecher and Palmer 2008). The goal of the postprocessing is then to obtain calibrated probabilistic forecasts. Two commonly used methods are nonhomogeneous Gaussian regression (Gneiting et al. 2005) and Bayesian model averaging (Raftery and Gneiting 2005).

Output from NWP models is typically made available for lead-time intervals between 1 and 6 h. It is then usually postprocessed separately for each lead time and location (weather station), which can become computationally expensive when performed for many stations and/or lead times. Scheuerer and Bürmann (2014) proposed a method to simultaneously postprocess...
spatially using anomalies. Stauffer et al. (2017) and Dabernig et al. (2017) took this method a step further by standardizing the anomalies. Their basic idea is to fit a statistical model on standardized anomalies (Wilks 2011) instead of the direct observations and ensemble forecasts. Since these standardized anomalies should not have any station-specific characteristics anymore, one statistical model can be used for all locations within a certain region.

We adapt the method of Dabernig et al. (2017), but instead of forecasting several stations simultaneously, we forecast several lead times simultaneously, which is computationally more efficient. Standardized anomalies remove lead-time-specific and seasonal characteristics of the data. Conventional methods are usually trained on a sliding window (Gneiting et al. 2005) to account for seasonal differences in the relationship between forecasts and observations. Standardized anomalies do not contain these season-specific characteristics anymore, so longer training datasets can be used. The implicit assumption is, however, that the NWP forecast skill does not vary seasonally. It will be shown that the third advantage of this approach is to provide calibrated forecasts at a much higher temporal resolution than the NWP forecasts by using a high-resolution observation climatology.

The remainder of this article is structured as follows: the data are described in the next section, followed by the methodology. Subsequently, results for a single station and for multiple stations are presented with a discussion and the conclusions completing this article.

2. Data

The proposed method is tested for the topographically complex terrain of South Tyrol in northern Italy from 1 February 2010 to 31 January 2016. Local time in this region is 1 h ahead of UTC. The forecast variable \( y \) is the temperature measured 2 m above ground. Ten-minute averages are available at 39 weather station locations. The predictor variable is the 2-m temperature of the ensemble NWP model from the European Centre for Medium-Range Weather Forecasts (ECMWF). The model has 51 members. Data are bilinearly interpolated from the 32-km model grid (T639) to the measurement sites. Lead times from the 0000 UTC run between +12 and +120 h are used with an interval of 3 h.

3. Method

a. Nonhomogeneous Gaussian regression

Gneiting et al. (2005) introduced nonhomogeneous Gaussian regression (NGR) for statistically postprocessing of ensemble forecasts. The observations are assumed to follow a normal distribution with regression equations for both the mean \( \mu \) and the standard deviation \( \sigma \):

\[
y \sim N(\mu, \sigma),
\]

\[
\mu = b_0 + b_1 m,
\]

\[
\log(\sigma) = c_0 + c_1 \log(s),
\]

with regression coefficients \( b_0, b_1, c_0 \) and \( c_1 \). Here \( \mu \) depends on the ensemble mean \( m \), and \( \sigma \) depends on the ensemble spread \( s \). We adapted the method of Gneiting et al. (2005) slightly by using the logarithmic link function in Eq. (3) to ensure positive values for the standard deviation \( \sigma \). Another adaption is fitting Eq. (1) by optimizing the maximum likelihood (with the R package cro; Messner et al. 2016) instead of minimizing the continuous ranked probability score (CRPS; Hersbach 2000; Gneiting and Raftery 2007).

b. Standardized anomalies

The basic idea of using standardized anomalies is visualized in Fig. 1. While the annual cycles clearly differ for the two different lead times and times within the diurnal cycle (top panel), the standardized anomalies for both lead times are centered around zero and have no pronounced annual cycle anymore (bottom panel). This may allow the use of the same coefficients for both lead times and to fit only one statistical model.

To obtain standardized anomalies, the climatological mean \( \mu_\text{y} \) is subtracted from the observations and divided by a climatological standard deviation \( \sigma_\text{y} \):

\[
\hat{y} = \frac{y - \mu_\text{y}}{\sigma_\text{y}}.
\]

The standardized anomalies \( \hat{m} \) and the corresponding \( \hat{s} \) are calculated as follows:

\[
\hat{m} = \text{mean}\left(\frac{\text{ens} - \mu_\text{ens}}{\sigma_\text{ens}}\right)
\]

and

\[
\hat{s} = \text{std dev}\left(\frac{\text{ens} - \mu_\text{ens}}{\sigma_\text{ens}}\right),
\]

where \( \text{ens} \) is the 51-member ensemble forecast and \( \mu_\text{ens} \) and \( \sigma_\text{ens} \) are the climatological mean and standard deviation, respectively. The \( \mu_\text{ens} \) is the same for all members. The \( \hat{m} \) is then the mean of all 51 standardized ensemble member anomalies and \( \hat{s} \) is their standard deviation. The details for computing \( \mu_\text{y}, \mu_\text{ens}, \sigma_\text{y}, \) and \( \sigma_\text{ens} \) will be shown in section 3d.
c. Standardized anomaly model output statistics (SAMOS)

Standardized anomalies remove seasonal and lead-time-specific characteristics and as a result several lead times of the same station can be fitted and forecast together, if the climatology is able to capture these different characteristics. Therefore, Eq. (1) is fitted with standardized anomalies ($\tilde{y}$, $\tilde{m}$, and $\tilde{s}$) of the observations and predictions instead of the observations and predictions themselves. Since the coefficients ($b_0$, $b_1$, $c_0$, and $c_1$) are fitted on data without lead-time specifics they are representative for every lead time. One underlying assumption is that the systematic error of the forecast uncertainty provided by the spread of the NWP ensemble does not change with lead time.

Once the coefficients are determined, a rearranged form of Eqs. (2) and (3) can be used for forecasting with new data:

$$\hat{\mu} = (b_0 + b_1 \tilde{m}) \times \sigma_\gamma + \mu_\gamma$$  \hspace{1cm} (7)

and

$$\hat{\sigma} = \exp[c_0 + c_1 \log(\tilde{s})] \times \sigma_\gamma.$$  \hspace{1cm} (8)

Since the coefficients do not depend on lead time, forecasts are no longer restricted to the lead times of the NWP forecasts. And importantly, the higher temporal resolution (e.g., 10 min) of the observation climatology ($\tilde{m}$ and $\tilde{s}$) provides additional information about the diurnal cycle in between lead times. Temporal information for individual events of abrupt temperature changes such as frontal passages, however, is solely provided by NWP data.

d. Climatology

To obtain forecasts with a high temporal resolution, a temporally highly resolved climatology is necessary. It can be produced with Generalized Additive Models for Location, Scale, and Shape (GAMLSS; Rigby and Stasinopoulos 2005; Stasinopoulos and Rigby 2007) as in Dabernig et al. (2017). With GAMLSS one model for

---

1 Because of the centering of the standardized anomalies $b_0$ is not significant and could be left out.
climatological mean and standard deviation for all lead times and all points in between can be calculated simultaneously. GAMLSS is similar to NGR but can also include nonlinear effects:  
\[ y \sim N(\mu_y, \sigma_y). \]  

To calculate a climatology, the daily variation has to be considered as well as the seasonal change. The nonlinear models for \( \mu_y \) and \( \sigma_y \) are therefore as follows:  
\[ \mu_y = \beta_0 + f(yd, dm) \quad \text{and} \quad \sigma_y = \gamma_0 + g(yd, dm), \]  

with \( \beta_0 \) and \( \gamma_0 \) as regression coefficients and \( f \) and \( g \) as smooth functions that capture the interaction between the day of the year (\( yd \)) and the minute of the day (\( dm, 0–1439 \)) with a cyclic cubic regression spline (fitted with the R package \texttt{gamlss}). Eight degrees of freedom are used in the function for the day of the year, and 10 degrees of freedom are used for the minute of the day.

The resulting terms (“effects”) of the fitted climatological mean based on 10-min measurements are illustrated in Fig. 2. Figure 2a shows the seasonal effect exemplified by four different times of the day indicated by the colored lines. The seasonal amplitude for the morning hours (bright colors) is smaller than in the afternoon (darker colors) and the shape and date of the peak also vary.

Figure 2b shows the diurnal effect exemplified by four different months. The minimum temperature occurs around 0900 UTC in January (brightest line) and about 3 h earlier in July (darkest line). These figures show the need for modeling an interaction term between seasonal and diurnal effects. The GAMLSS model provides then a climatology for every day of the year and every minute of the day.

In addition to climatologies of the observations, climatologies of the ensemble forecasts are also required, which can be computed analogously. However, since forecasts in our case are only available every 3 h instead of every 10 min, we model only a seasonal effect separately for each lead time and interpolate linearly in between. Equation (10) is replaced by  
\[ \mu_{\text{ens}} = \beta_0 + f(yd) \quad \text{and} \quad \sigma_{\text{ens}} = \gamma_0 + g(yd), \]  

where nonlinear effect for the day of the year uses eight degrees of freedom.

4. Different models

Three different models with standardized anomalies are compared against a reference to investigate the advantages of SAMOS. The reference is an NGR [ensemble model output statistics (EMOS); Gneiting et al. 2005] fitted separately for each 3-hourly lead time with a sliding window of the previous 30 days as training data. The SAMOS variations differ in the calculation of the observation climatology and NGR. The observation climatology can either be calculated at every lead time individually as in Eq. (11) or at all lead times simultaneously as in Eq. (10). As second difference, Eqs. (7) and (8) can either be calculated separately for each lead time or simultaneously for all lead times. The three SAMOS variations are as follows:

- **SAMOS lead-time-wise (SAMOS ltw)**, where the climatologies and the NGR equations are calculated separately for each 3-hourly lead time. This model shows the effect of employing standardized anomalies.
- **For the second model SAMOS lead-time-wise simultaneous (SAMOS ltw-S)** the climatology to compute
the standardized anomalies is also fitted separately for all lead times, but one single NGR equation is calculated simultaneously for all lead times.

- The third model is SAMOS full as described in the methods section with one simultaneous climatology over all available observations and one simultaneous NGR for all 3-hourly lead times.

A forecast example of SAMOS full for a particular initial day is shown in Fig. 3. Whereas conventional methods such as EMOS only produce forecasts at the 3-hourly lead times, SAMOS full also provides forecasts in between. The forecasts in between the 3-hourly lead times follow the shape of the observation climatology (dashed line in Fig. 3). Additionally, SAMOS fully inflates the uncertainty of the raw NWP ensemble to correct for its underdispersion (Hopson 2014; Möller and Groß 2016; Dabernig et al. 2017).

All models are trained on data independent from that used for verification. For the reference model EMOS this is achieved by training it with data of a rolling window from the 30 days prior to the forecast day and verifying it with the remaining days. For the SAMOS variants this is achieved by tenfold cross validation. The complete dataset of 6 yr is split into 10 consecutive parts. Each tenth is then forecast and verified with a model trained on the remaining nine parts.

5. Results

The performance of these forecasts is evaluated using different metrics. The mean absolute error (MAE) is used as deterministic measure while the CRPS is used to evaluate the probabilistic performance. Both are calculated separately for each lead time and model with tenfold cross-validated data. The metrics are shown as skill score relative to a reference model (ref):

$$\text{skill score} = 1 - \frac{\text{score}}{\text{score}_{\text{ref}}}.$$  \hfill (12)

To investigate calibration, the root-mean-squared error (RMSE) of the predicted mean $\bar{m}$ is compared to the predicted standard deviation $\bar{s}$.

In the following, we first show results for a single station (Auer) before aggregating results of all 39 stations.

a. Single station

To evaluate how well the models are calibrated, the RMSE and the standard deviation of the forecasts are compared in Fig. 4. A well-calibrated forecast should have a predicted standard deviation similar to the RMSE of the predicted mean. Figure 4a shows that the standard deviation of the reference EMOS model is less than the RMSE of the mean and the model at all lead times and is thus underdispersive. Standard deviation and RMSE of SAMOS lead-time-wise (Fig. 4b), on the other hand, are almost identical, showing the model to be almost perfectly calibrated. When the NGR equations are no longer fitted separately but simultaneously for all lead times (dotted line in Fig. 4b) results remain similar (RMSE not significantly different to SAMOS lead-time-wise) except for an overdispersion in the afternoon of the first four forecast days. This remains unchanged when the climatologies are also fitted simultaneously (SAMOS full, Fig. 4c).

The results for the deterministic MAE and the probabilistic CRPS metrics are illustrated in Fig. 5. Forecasts in between the 3-hourly lead times for EMOS, SAMOS lead-time-wise, and SAMOS lead-time-wise simultaneous are interpolated with a natural spline function to the 10-min resolution.
spline was used instead of a linear interpolation to better capture the diurnal cycle of the temperature. Forecast quality varies with the time of the day. It is best (lowest values) in the second half of the night and the morning (Fig. 5a). This pattern is already present in the raw ECMWF forecasts (not shown). The three SAMOS models outperform the EMOS reference model most of the time with the largest improvement in the afternoon.

The effect of fitting climatologies and NGR equations simultaneously (SAMOS full) instead of separately for each lead time (SAMOS-ltw) on the forecast quality is shown in Fig. 5b by the CRPS skill score with SAMOS-ltw as reference. The 90% confidence interval of the mean CRPS skill score is computed from 500 bootstrap samples and shown as a gray area in Fig. 5b. On average, the forecasts of SAMOS full are half a percent worse at the 3-hourly lead times (red dots) but similar or better in between with the biggest improvements farthest from the lead times.

b. All stations

So far we have only evaluated the performance at a single forecast location (Auer). To obtain more general results, the individual results of the 39 stations in the region of South Tyrol will be combined. The MAE and CRPS skill scores with EMOS as reference are averaged over all stations for each lead time. One set of skill scores is computed at the 3-hourly resolution of the NWP model; the other at the 10-min resolution of the observations. The results are shown in Fig. 6.

The deterministic MAE at the 3-hourly lead times, for which NWP input is available, is slightly but not significantly better than the EMOS reference for SAMOS lead-time-wise and SAMOS lead-time-wise simultaneous, and slightly worse for SAMOS full. In contrast, the probabilistic CRPS metric of all SAMOS models on the 3-hourly lead times is significantly better than that of EMOS. This difference between MAE and CRPS shows that the SAMOS models capture the forecast quality variations more accurately than the EMOS model.
uncertainty better than EMOS, which can also be seen in Fig. 4. SAMOS full is still slightly worse than the other SAMOS variations.

At the 10-minute resolution, however, SAMOS full performs as well as the other SAMOS variants. The better performance of SAMOS full in between the 3-hourly ECMWF lead times compensate for the slightly worse performance at these lead times.

6. Discussion and conclusions

Dabernig et al. (2017) and Stauffer et al. (2017) showed that postprocessing NWP ensemble forecasts with standardized anomalies instead of the forecast variable itself allows the use of one single statistical model for forecasts at an arbitrary location. The traditional approach of ensemble model output statistics (EMOS; Gneiting et al. 2005) thus becomes standardized anomaly model output statistics (SAMOS). The current paper takes this approach from the spatial into the temporal domain and fits one single statistical model to several lead times. This improves the forecast skill, increases the temporal resolution of the forecasts, and reduces the time needed to compute calibrated probabilistic forecasts. The concept is demonstrated for forecasts of 2-m temperature in the region of South Tyrol in Italy. The topographic complexity of this region and the ensuing large differences between real and NWP model topography create larger systematic NWP model errors than over flat terrain. In addition, these systematic model errors vary diurnally and seasonally.

Standardized anomalies remove diurnal and seasonal cycles from the data. Consequently, all available data can be used to fit the postprocessing statistical models. Season-specific characteristics have also been handled by using a moving training window of subseasonal width—typically 30 days (e.g., Gneiting et al. 2005). Being able to use all data instead of only data within the window makes the fit more stable and increases the forecast skill. In addition, the statistical model is formulated over a wider range of weather events and covers more of climatologically possible scenarios. The training data could be expanded even further and the forecast skill improved even further with reforecasts as shown by Stauffer et al. (2017). Even then events more extreme than contained in typical reforecast periods of 20–30 years can occur. As long as the NWP model is capable of producing forecasts that are extreme compared to the NWP climatology, SAMOS should be able to produce forecasts that are extreme compared to the observation climatology.

The underlying assumption for using standardized anomaly data over all seasons is that systematic errors in the NWP ensemble spread have no strong seasonal
dependence. If that assumption were violated, an additional seasonal term in the NGR equation [Eq. (3)] could be used to accommodate this dependence.

Both observations and NWP model output can be transformed to standardized model anomalies by subtracting the respective means and dividing by the respective standard deviations. Merely replacing the direct values, which are used in the EMOS post-processing approach with standardized anomalies in the SAMOS approach (termed SAMOS lead-time-wise) improves the mean absolute error (MAE) as deterministic metric and the continuous ranked probability score (CRPS) as probabilistic measure. SAMOS also improves the calibration and reduces the underdispersion of the NWP model output even more than EMOS.

When the nonhomogeneous regression models are no longer fit separately for each lead time but simultaneously for all lead times (SAMOS lead-time-wise simultaneous) the verification metrics are similar. Computational speed is increased and the complexity of running a probabilistic postprocessing system in an operational setting with hundreds or thousands of locations is reduced. SAMOS full forecasts are an order of magnitude faster to compute than EMOS over a simulated year of operational forecasts.

Finally, when the statistical models for mean and standard deviations of observations and NWP model output needed for the computation of standardized anomalies are also fit simultaneously for all times (SAMOS full), the temporal resolution of the forecasts can become as high as the temporal resolution of the observations. Output of ensemble NWP models is typically available every few hours (3 in our case), while measurements are typically available several times per hour (every 10 min in our case). The increase in temporal resolution is not simply an interpolation but provides additional information contained in the highly resolved seasonally varying daily cycle of the observations. Abrupt temperature changes from particular synoptic or mesoscale events without a pronounced diurnal cycle such as a frontal passage are smoothed out in the climatology of the observations. The arrival time of such an event can therefore not be more finely resolved than in the NWP model. Nor can a possible timing error of such an event be corrected. Multivariate methods such as ensemble copula coupling (Schefzik et al. 2013) would be needed for such a correction.

An alternative approach to achieving a temporal resolution of the forecasts as high as the resolution of the observations is to first interpolate the raw ensemble NWP model output to observation times and then fit a statistical model separately for each lead time. Five-day
forecasts at a 10-min resolution would then require 649 separate models per station, which is impractical.

Acknowledgments. We thank two anonymous reviewers for their valuable comments that improved this manuscript substantially. The authors also thank the Austrian weather service ZAMG for providing access to ECMWF data and the South Tyrolean weather service for the observations. This study is supported by the Autonome Provinz Bozen (Abteilung Bildungsförderung, Universität und Forschung—ORBZ110725).

REFERENCES

Dabernig, M., G. J. Mayr, J. W. Messner, and A. Zeileis, 2017: Spatial ensemble post-processing with standardized anomalies. Quart. J. Roy. Meteor. Soc., 143, 909–916, doi:10.1002/qj.2975.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteor., 11, 1203–1211, doi:10.1175/1520-0450(1972)011...TUOMOS>2.0.CO;2.

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

Gneiting, T., A. H. Westveld, and T. Golfman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Wea. Rev., 133, 1098–1118, doi:10.1175/MWR2904.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. Wea. Forecasting, 15, 559–570, doi:10.1175/1520-0434(2000)015...DOTCRP>2.0.CO;2.

Hopson, T. M., 2014: Assessing the ensemble spread–error relationship. Mon. Wea. Rev., 142, 1125–1142, doi:10.1175/MWR-D-12-00111.1.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. J. Comput. Phys., 227, 3515–3539, doi:10.1016/j.jcp.2007.02.014.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. Tellus, 34A, 505–513, doi:10.1111/j.2153-3490.1982.tb01839.x.

Messner, J. W., G. J. Mayr, and A. Zeileis, 2016: Heteroscedastic censored and truncated regression with crch. R J., 8 (1), 173–181.

Möller, A., and J. Groß, 2016: Probabilistic temperature forecasting based on an ensemble autoregressive modification. Quart. J. Roy. Meteor. Soc., 142, 1385–1394, doi:10.1002/qj.2741.

Raftery, A., and T. Gneiting, 2005: Using Bayesian model averaging to calibrate forecast ensembles. Mon. Wea. Rev., 133, 1155–1174, doi:10.1175/MWR2906.1.

Rigby, R. A., and D. M. Stasinopoulos, 2005: Generalized additive models for location, scale and shape. J. Roy. Stat. Soc. Ser. C Appl. Stat., 54, 507–554, doi:10.1111/j.1467-9876.2005.00510.x.

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

Scheuerer, M., and L. Bürmann, 2014: Spatially adaptive post-processing of ensemble forecasts for temperature. J. Roy. Stat. Soc. Ser. C Appl. Stat., 63, 405–422, doi:10.1111/rasc.12040.

Stasinopoulos, D. M., and R. A. Rigby, 2007: Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. J. Stat. Software, 23, 1–46, doi:10.18637/jss.v023.i07.

Stauffer, R., N. Umlauf, J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Ensemble postprocessing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. Mon. Wea. Rev., 145, 955–969, doi:10.1175/MWR-D-16-0260.1.

Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences. Vol. 100. 3rd ed. Academic Press, 704 pp.