Operational statistical postprocessing of temperature ensemble forecasts with station-specific predictors

Kaisa Ylinen | Olle Räty | Marko Laine

Abstract
A proper account for forecast uncertainty is crucial in operational weather services and weather-related decision-making. Ensemble forecasts provide such information. However, they may be biased and tend to be under-dispersive. Therefore, ensemble forecasts need to be post-processed before using them in operational weather products. The present study post-processes the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system temperature forecasts over Europe with lead times up to 240 hr using the statistical calibration method that is currently implemented in operational workflow at the Finnish Meteorological Institute (FMI). The calibration coefficients are estimated simultaneously for all stations using a 30 day rolling training period. Station-specific characteristic are accounted for by using elevation, latitude and land–sea mask as additional predictors in the calibration. On average the calibration improved the ensemble spread over Europe, although the improvements varied between different verification months. In March, the calibration improved ensemble forecasts the most, while in January the performance depended strongly on location. A comparison between three versions with different sets of station-specific predictors in calibration showed that elevation was the most important predictor, while latitude and land–sea mask improved the forecasts mostly with shorter lead times. The calibration for the Finnish stations was also tested using three different size training domains in order to find the optimal training area. The results showed that smaller training domains had a significant effect on calibration performance only at lead times up to a few days. With longer lead times, the calibrated forecasts were better when all available stations were included.

KEYWORDS
non-homogeneous Gaussian regression, probabilistic forecasting, station specific, statistical ensemble post-processing, temperature, weather forecasting

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited. © 2020 The Authors. Meteorological Applications published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.
1 | INTRODUCTION

Numerical weather prediction (NWP) models form the basis of many weather forecast products used by the public audience and stakeholders for weather forecast guidance. Different NWP models support weather forecasting from daily to weekly forecast scales and also from monthly to seasonal scales (e.g. Molteni et al., 1996; Willett et al., 2008). Deterministic forecasts are used as the basis of many operational forecast products. However, their usability is limited in some cases due to the following issues. While NWP models can resolve most of the large-scale physical processes in the atmosphere sufficiently, some processes are totally excluded or represented inadequately. As the initial analysis also has errors, deterministic forecasts are usually biased in comparison with observations. Even more importantly, deterministic forecasts do not provide any information on forecast uncertainty. Rather than integrating a single deterministic forecast from a supposedly best guess of the initial state, it has been shown that a better approach would be to make ensemble predictions using a set of forecasts with slightly different initial conditions or different formulations of a forecast model (Palmer, 2000). The probabilistic output of an ensemble forecast system is a useful tool for end-users to base the decisions on, for example, the impacts of hazardous extreme weather events.

Ideally, the ensemble spread should be equal to the true probability distribution of the forecast uncertainty. In reality, medium-range ensemble forecasts tend to have typically too high forecast confidence compared with actual forecast uncertainty, and they may also suffer from local forecast biases. Post-processing of ensemble forecast output has been used to improve forecast reliability and accuracy. Most of the recently used statistical methods share the same general approach of correcting the current forecast by using the past forecast errors, as has been done for deterministic forecasts in the so-called model output statistics (MOS) procedure introduced originally by Glahn and Lowry (1972). This process makes use of information from the historical forecasts and observations to produce probabilistic forecasts and to improve their reliability. First operational ensemble forecasting systems were developed in the 1990s, and NWP models have improved a lot since then. There is still a need for calibration, as model development and ensemble calibration concentrate on different aspects of forecast improvement (Hemri et al., 2014). The calibration of ensemble forecasts has been shown to be useful at a variety of forecast time scales, including lead times from short (e.g. Raftery et al., 2005; Feldmann et al., 2015) to medium range (e.g. Hagedorn et al., 2008).

Several different statistical methods and approaches have been developed for post-processing purposes (e.g. Hansen and Emanuel, 2003; Wilks, 2015; Rosgaard et al., 2016), the method providing the best outcome depending on the weather variable being forecasted. Non-homogeneous Gaussian regression (NGR) is well known and widely used to calibrate ensemble weather forecasts (Raftery et al., 2005). This parametric method is able to represent and correct both bias and dispersion errors while at the same time being relatively simple in formulation. The model parameters can be estimated locally at each observation point or jointly over a region (Gneiting et al., 2005; Thorarinsdottir and Gneiting, 2010). As forecast biases vary regionally, partly due the coarse resolution of model orography and land–sea distribution, the local approach typically has a better predictive performance than the regional approach. However, local calibration allows predictions only at observation locations and needs more historical data to estimate the calibration coefficients robustly. From an operational perspective, regional coefficient estimation requires fewer computation resources and provides statistically post-processed forecast for spatially continuous domains.

These two calibration strategies can be combined to exploit both their strengths. To handle local biases better, we used elevation, latitude and land–sea mask (lsm) as station-specific predictors in the NGR model in addition to ensemble mean and standard deviation. Other spatially adaptive approaches have been studied for temperature and wind speed forecasts. A similar idea as that here is used in cluster-based approaches, where the calibrated stations are grouped by their altitude (Diaz et al., 2019) or climatological characteristics (Lerch and Baran, 2017). Instead of grouping the stations, we try to include the station-specific characteristics linearly to the model. Also, more complex methods have been studied that combine the NGR with spatially dependent climatological anomaly fields (Feldmann et al., 2015; Möller et al., 2015; Dabernig et al., 2017).

A variant of the NGR method, with elevation being added to the model as a station-specific predictor, was used to calibrate ensemble forecasts from the Grand Limited Area Model Ensemble Prediction System (GLAMEPS). The GLAMEPS (Iversen et al., 2011) was run up to a 54 hr lead time in the whole European domain using the European Centre for Medium-Range Weather Forecasts ensemble forecasts (ECMWF-ENS) as the boundary model. The model was developed by the HIRLAM-C consortium, which consists of several European meteorological institutes. The operational GLAMEPS runs ended in July 2019, and nowadays the development work within the consortium is concentrated on other operational limited area models (Frogner et al., 2019).

The same method used to calibrate the GLAMEPS model was implemented and recently added to the operational workflow at the Finnish Meteorological Institute (FMI) with the initial focus being on the ECMWF-ENS 2 m temperature forecasts. Here we describe the method,
its implementation details and the general calibration approach used in operational work. We comprehensively evaluate its skill from an operational perspective and demonstrate the temporal and spatial variations in the performance of the calibration method at all lead times up to 10 days. However, typically for the short lead times, ensemble forecasts are provided by regional high-resolution NWP models, hence the main focus of the ECMWF-ENS forecasts is for lead times between 3 and 10 days. The present study uses forecasts initialized in September 2018 and January, March and June 2019 to investigate the performance of the calibration method in all seasons. We also test the use of latitude and lsm as additional station-specific predictors in addition to elevation in the NGR model to ascertain whether they could improve the calibration results. To cover the Finnish national weather service perspective, we inspect the calibration results in more detail over Finland. We also examine whether it is better to use all available stations or only those from a limited area for training the method to obtain regionally optimal calibration results, using Finland as the target region.

The remainder of the paper is structured as follows. Section 2 describes the observation and forecast data sets used. The calibration method and verification metrics for evaluating the ensemble forecast skill are presented in Section 3. Section 4 reports the verification results of the calibrated forecasts. Finally, conclusions are provided in Section 5.

2 | FORECAST AND OBSERVATION DATA

The ECMWF’s operational ensemble forecast system (ENS) consists of 51 members; it is run twice a day with 0000 and 1200 UTC initial times (Molteni et al., 1996; Buizza et al., 2007). The temporal resolution of the forecasts is 3 hr for lead times up to 144 hr, and 6 hr for lead times from 150 to 360 hr. The present study used 0000 UTC forecasts up to 240 hr. The ECMWF model covers the entire globe, with a current horizontal resolution of about 18 km. As the study's focus is on the European region, forecast runs were extracted on a 0.2° longitude × 0.2° latitude grid from 32.60 to 73.40° N and from 26.00° W to 42.40° E (Figure 1). For calibration coefficient estimation, 2 m temperature forecasts were interpolated bilinearly to the observation station points, and elevation was corrected with a static lapse rate of 6.5 K-km⁻¹ using the difference between the ECMWF-ENS model topography and the true elevation of the meteorological station to present the temperature at the altitude of the observation station. Temperature observations were gathered from European surface synoptic observations (SYNOP) stations. The locations of these about 2,700 stations that were operationally available for calibration are shown in Figure 1.

3 | CALIBRATION AND VERIFICATION METHODS

3.1 | Non-homogeneous Gaussian regression (NGR)

For the calibration of 2 m temperature forecasts, we used an NGR (e.g. Gneiting et al., 2005), which generalizes the common MOS post-processing technique. Calibrated temperature forecasts follow Gaussian distribution $N$:

$$N(a + b_{\text{ENS}}, c + dS^2),$$

FIGURE 1: Training data domains (1, Europe; 2, Northern Europe; and 3, Fennoscandia) and station locations (dots) for the calibration of European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble forecasts. The number of stations in each domain is shown in parenthesis. Finland, which is the focus region of this study, is shaded.
where $\overline{\text{ens}}$ is the ensemble mean; $S$ is the ensemble standard deviation; and $a$–$d$ are regression coefficients. This is the simplest form of the NGR where all the predictors are related to the ensemble members. This form is suitable if the NGR is estimated at every station individually. Another option is to fit calibration coefficients for all stations simultaneously. It has advantages in operational systems, since the coefficients must be estimated only once, and less historical data are required. In addition, the calibration can be applied to any location within the domain of interest, for example, on a regular grid or a series of locations of interest. Hence, forecasts are directly available in all locations with spatial coherence. Although the systematic errors in the model might have station-specific characteristics, they can be eliminated by adding station-specific predictors to the NGR model. Such a predictor in the calibration of the GLAMEPS model was elevation, which is also used in the operational version of the calibration method at the FMI for the ECMWF-ENS temperature forecasts. In this version, the parameters of the Gaussian distribution ($\mu$, $\sigma$) are expressed by the following equations:

$$
\begin{align*}
\mu_{kl} &= a_l + b_l \overline{\text{ens}}_{kl} + c_l \text{alt}_k, \\
\log(\sigma_{kl}) &= d_l + e_l \log(\text{sd}_{kl}) + f_l \log(\max(1, \text{alt}_k)),
\end{align*}
$$

(2)

where $\overline{\text{ens}}_{kl}$ is the mean and $\text{sd}_{kl}$ is the standard deviation of the 51 ensemble members for each location $k$ and lead time $l$; and the variable $\text{alt}_k$ indicates the average elevation (m above mean sea level) for each location $k$. In the operational model, alt is interpolated bilinearly from the ECMWF-ENS topographical field. For $\sigma$, the logarithmic link is used to ensure that the scale parameter is positive. In the following this version of the NGR is called $\text{cal\_oper}$.

The regression coefficients $a_l$–$f_l$ were estimated using point observations and forecasts from the 30 most recent days in a sliding-window manner. The coefficients were estimated separately for each lead time $l$ and forecast cycle (0000 UTC in the present study). Furthermore, to reduce the need for computation, the regression coefficients were updated once a week. For the calibration, we used the R-package $\text{gamls}$, which provides maximum likelihood estimates for the regression coefficients (Rigby and Stasinopoulos, 2005).

When estimating the NGR model, it would be best to have multiple years of historical training data (Lang et al., 2015; Möller et al., 2015; Diaz et al., 2019). To test the sensitivity of coefficient calibration to the length of the training period, we compared 15, 30 and 90 day sliding windows when estimating the calibration coefficients. The 15 day window had the best performance at shorter lead times, while the 90 day window provided the best results with the longest lead times. Still, the 30 days provided the best results on average over all lead times and was selected for the coefficient calculation.

The calibration coefficients were estimated using observations from the SYNOP stations and forecasts that are bilinearly interpolated to the station points. We used only the interpolated forecasts and performed both calibration and verification at the station locations. Note that in this case the location index $k$ stands for the actual station location. Because the calibration coefficients are global, the calibration is applied directly to gridded forecasts in the operational system at the FMI, and all the station-specific variables have to be available at all grid points.

The NGR predictive distribution was created for each location and lead time, and then converted back to ensemble members by extracting evenly spaced quantiles from the calibrated distribution. The rank of each member was maintained by reordering the members to the same order as in the raw ensemble to restore the spatio-temporal correlation structure of the ensemble. This is essentially the same as ensemble copula coupling with equidistant quantiles (Schefzik et al., 2013). The dependence between the ensemble members is essential if the ensemble forecasts are used as an input for downstream applications such as the forest fire weather index (Van Wagner, 1987) or renewable energy applications (Pinson and Messner, 2018).

As a simple reference model, the NGR is applied without any station-specific predictors ($\text{cal\_basic}$):

$$
\begin{align*}
\mu_{kl} &= a_l + b_l \overline{\text{ens}}_{kl}, \\
\log(\sigma_{kl}) &= c_l + d_l \log(\text{sd}_{kl}).
\end{align*}
$$

(3)

Thus, the reference model takes into account the effect of the ensemble mean and standard deviation in the calibration only. In addition to elevation, we tested latitude and lsm as additional station-specific predictors in the NGR model ($\text{cal\_full}$). Latitude could explain the bias variations between northern and southern areas of Europe. The lsm is a potential predictor to modify biases that appear near sea areas when the sea is either warmer or cooler than the ambient air. All these additional station-specific predictors were important in the calibration model when judged by their $p$-values. Also, interaction terms between the ensemble mean and station-
specific predictors are included in the model since they have a significant effect on the outcomes. For more details, see Tables S1–S4 in Appendix S1 in the additional supporting information, where the fraction of training cases with \( p < 0.05 \) is shown for each predictor.

With all station-specific parameters included, the NGR has the following formulas for \( \mu \) and \( \sigma \):

\[
\begin{align*}
\mu_{kl} &= a_l + b_l \text{lsm}_{kl} + c_l \text{alt}_k + d_l \text{lat}_k + e_l \text{lsm}_k \\
&\quad + f_l \text{lsm}_{kl} \text{alt}_k + g_l \text{lsm}_{kl} \text{lat}_k + h_l \text{lsm}_{kl} \text{lsm}_k, \\
\log(\sigma_{kl}) &= i_l + j_l \log(\text{sd}_{kl}) + k_l \text{lsm}_{kl} + l_l \text{alt}_k + m_l \text{lat}_k \\
&\quad + n_l \text{lsm}_k.
\end{align*}
\]  

Since the lsm is related to the model grid only, the lsm values were interpolated (nearest neighbour) before coefficient estimation to each station location from the ECMWF-ENS lsm field, which ranges from 0 to 1.

### 3.2 Verification metrics

The aim of the ensemble calibration is to maximize the sharpness of the ensemble forecasts subject to calibration that refers to reliability (Gneiting et al., 2007). The predictive performance of ensemble forecasts can be assessed with multiple verification tools (Wilks, 2011). A common way to assess forecast reliability is to compare the standard deviation of an ensemble with the average root mean square error (RMSE) of the ensemble mean. The spread should be equal to the RMSE on average in a well-calibrated ensemble.

The overall forecast performance can be assessed via proper scoring rules, which are able to judge calibration and sharpness simultaneously. A widely used proper scoring rule is the continuous ranked probability score (CRPS), which compares the full distribution of an ensemble forecast with the observed target, where ensemble forecasts are represented as cumulative distribution functions (CDF) (Gneiting and Raftery, 2007). The CRPS is calculated using:

\[
\text{CRPS} = \int_{-\infty}^{\infty} (F_F(x) - F_o(x))^2 \, dx, \tag{5}
\]

where \( F_F(x) \) is the forecast CDF; and \( F_o(x) \) is the observed CDF. The CRPS is expressed in the same units as the observation, with smaller values indicating a better performance for the ensemble forecasts. The present study used the analytical Gaussian form of the CRPS for both raw and calibrated ensemble forecasts.

To evaluate the actual skill against raw ensemble forecasts, a continuous ranked probability skill score (CRPSS) was calculated using:

\[
\text{CRPSS} = 1 - \frac{\text{CRPS}_{\text{cal}}}{\text{CRPS}_{\text{ref}}}, \tag{6}
\]

where subscripts cal and ref denote the CRPS of calibrated and raw ensemble forecasts, respectively. All the statistical analyses were conducted with the statistical software R (R Core Team, 2019).

### 4 RESULTS

This section presents the results of applying the introduced calibration methods to ECMWF-ENS forecasts. First, we introduce the general performance of the operational version of the NGR (\textit{cal-oper}) over the whole of Europe, together with a regional inspection over Finland. Next, we compare the verification results for the calibrated forecasts obtained with and without different sets of station-specific predictors. Lastly, the results for Finnish stations are evaluated in more detail, and the effect of using different training domain sizes is assessed.

#### 4.1 Operational NGR method

We begin by evaluating the operational NGR method (\textit{cal-oper}) over all stations in Europe. Figure 2 shows the spread-skill results for both the raw and calibrated ECMWF-ENS forecasts in the four study months. For raw forecasts, the spread (dashed line) is constantly smaller than the forecast error (solid line), especially over the first few days of the medium-range forecasts. After calibration, the ensemble spread is systematically closer to the forecast error, which indicates that the reliability of ensemble forecasts is improved. In all months except January, the spread is almost equal to the skill. Calibration also decreases the RMSEs slightly in most cases, which means that the accuracy of the ensemble forecasts is improved.

Besides evaluating \textit{cal-oper} with all the available stations in Europe, verification results were also analysed regionally in Finland. Figure 3 shows the spread-skill results of the raw and calibrated ECMWF-ENS forecasts averaged over all stations in Finland for four different months. The calibration seems to work best in June and also improves the results in March and September. However, in March the calibrated forecasts are still slightly under-dispersive, especially during the night, while in
FIGURE 2  Spread and skill (root mean square error—RMSE) for the raw and calibrated European Centre for Medium-Range Weather Forecasts ensemble forecasts (ECMWF-ENS) averaged over all stations in Europe: September 2018 and January, March and June 2019. Calibration was made with cal_oper

FIGURE 3  Spread and skill (root mean square error—RMSE) for the raw and calibrated European Centre for Medium-Range Weather Forecasts ensemble forecasts (ECMWF-ENS) averaged over all stations in Finland (number of stations = 142) and shown separately for all study months. Calibration was made with cal_oper using all available stations from the European domain
September calibration does not follow the diurnal cycle in the ensemble skill, leading to overdispersion during the day. In the January 2019 case, \textit{cal\_oper} fails to calibrate the ensemble forecasts in Finland. Even though the spread and skill relationship is slightly improved up to an 80 hr lead time, the RMSE is increased at the same time in comparison with the raw ensemble forecasts. The reason for the larger RMSE at all lead times is that the ECMWF model tends to over-forecast very low temperatures in northern Finland and Sweden (Figure 4), while the overall bias is mostly negative in other parts of Europe. This leads to increases in forecast biases in Finland after the calibration (Figure 5). It is well known that the ECMWF-ENS cannot capture very low temperatures due to model limitations (Hyrkkänen \textit{et al.}, 2016). For example, the ECMWF model forecasted in January 2019 always $-20^\circ$C at its lowest in northern Finland whether the real temperature was $-20$ or $-30^\circ$C. In other seasons, there is either no clear bias signal present or the bias is more uniform as is usually seen in the spring over Europe (Figure 6).

### 4.2 Effect of station-specific predictors

In all previous results, only elevation was used to provide station-specific information to the calibration. To understand better the effect of station-specific predictors on calibration performance, we next compare the operational version of our NGR model with two alternative versions. In the first (\textit{cal\_basic}), all station-specific predictors are omitted. The second alternative (\textit{cal\_full}), on the other hand, uses latitude and the lsm together with elevation as sources of local information to calibrate the forecast mean and spread.

To compare the three model versions, the CRPSS was calculated at every station in Europe using raw ensemble forecasts as the baseline. Figure 7 shows that, on average, all calibration methods result in positive CRPSS at short lead times (3–72 hr). However, in all cases there are some stations where calibration actually worsens the ensemble forecasts. The results also show that \textit{cal\_basic} improves the CRPSs of raw ensemble forecasts on average by 9.1% in September, 6.4% in January, 7.9% in March and 9.7%
FIGURE 6  Mean bias for the raw European Centre for Medium-Range Weather Forecasts ensemble forecasts (ECMWF-ENS) in March 2019 at each station with short, medium and long lead times.

FIGURE 7  The distribution of the continuous ranked probability skill score (CRPSS) at all stations averaged over lead times from 3 to 72 hr with three different calibration models: without station specific parameters (left), with elevation as an additional predictor (centre) and with elevation, latitude and land–sea mask (lsm) as additional predictors (right). The average CRPSS is shown above each violin plot. The reference model is the raw ensemble.
in June, while with cal_oper the improvements are still larger by 10.2%, 8.9%, 10.8% and 10.4% in the respective months. Using all three station-specific parameters (cal_full) further improves the calibrated forecasts, the improvement in the CRPSSs being 11.2% in September, 10.0% in January, 12.5% in March and 10.5% in June. Differences in the average CRPSSs between the methods are statistically significant (at the 5% significance level), except in June between cal_oper and cal_full (see Table S5 in Appendix S1 in the additional supporting information). When investigating the effect of each station-specific parameter individually in the NGR model, elevation had the largest positive effect in September, January and March based on the averaged CRPSSs (data not shown). In June, the effects of elevation and the lsm were equally large. Also, latitude and the lsm individually had a positive effect on the calibration model in all months compared with cal_basic.

With lead times from 75 to 144 hr (see Figure S1 in Appendix S1 in the additional supporting information) the differences in CRPSS between cal_basic and cal_oper are mostly similar to those seen at shorter lead times. However, the CRPSS for cal_full indicates that the benefit of using additional station-specific predictors has mostly vanished. In March and June, the differences between cal_oper and cal_full are not statistically significant (at the 5% significance level). As an individual predictor, elevation had the most important effect on the calibration model in all months except June when the lsm improved the model the most. With lead times from 150 to 240 hr (see Figure S2 in Appendix S1 in the additional supporting information), the differences between methods are even smaller; however, there are statistically significant difference (at the 5% significance level) in all cases except March between cal_basic and cal_full (see Table S5 in Appendix S1 in the additional supporting information). Also, there is no benefit of using the two additional station-specific predictors, while cal_oper still performs slightly better compared with cal_basic. Again, the elevation was the most important individual predictor in our calibration model in September, and in March both elevation and the lsm had an equally positive effect on the calibration model. In June, the lsm improved the model the most, while latitude was most important in January, but in September and March latitude deteriorated the calibration model compared with cal_basic. Given the deterioration of the CRPSS for cal_full at longer lead times, we concluded that cal_oper has the overall best performance. For the sake of completeness, spatial maps of the differences in the CRPSSs are illustrated in Figures S3–S6 in Appendix S1 in the additional supporting information.

4.3 Different training areas

Since forecast errors might differ substantially between southern and northern parts of Europe, we tested how the size of the training domain affects the calibration results at stations in Finland. Three training domains were considered: whole of Europe, northern half of Europe and Fennoscandia (Figure 1). All three calibration methods were included and trained using all stations at each of the three training domains. The CRPSSs calculated over lead times between 3 and 72 hr are shown for all stations in Finland in Figure 8. The CRPSSs are improved for Finnish stations when using either one of the smaller training domains, especially with cal_basic, which does not include any local predictors. The CRPSS is also somewhat improved for cal_oper when the two smaller training domains are used, but the improvement is less pronounced than for cal_basic.

In January, the improvement in the CRPSS is largest with smaller training domains for both cal_basic and cal_oper. The large improvement in the CRPSS for these two methods is understandable in the January 2019 case. The previously discussed location-dependent bias can be captured by the two simpler methods only by limiting the training domain to cover those stations in which the forecast bias behaves similarly. When the additional station-specific predictors are included (cal_full), all training domains provide comparable skill, as enough location-specific information on forecast bias is covered by these predictors. Overall, it seems that with short lead times, the use of additional station-specific predictors provides similar improvement for calibrated forecasts than using a smaller training domain, but with less predictors such as the linear latitude effect.

With lead times between 75 and 144 hr (see Figure S7 in Appendix S1 in the additional supporting information), using a smaller training area improves the CRPSSs only slightly, except in January, due to the strong location-dependence on model bias. With lead times between 150 and 240 hr (see Figure S8 in Appendix S1 in the additional supporting information), the smaller training areas do not improve the CRPSSs anymore, and in many cases the best CRPSSs are achieved using the whole of Europe as the training domain, especially in March. Figure S8 in Appendix S1 in the additional supporting information also shows that with lead times up to 150 hr, the calibration improves the forecasts in Finland only in March and June. As an overall conclusion, depending on the lead time and forecast case, there are no or only minor improvements in the CRPSSs when the smaller training domains are used.
This paper examined the calibration of the European Centre for Medium-Range Weather Forecasts ensemble forecasts (ECMWF-ENS) for 2 m temperature with lead times up to 240 hr using the non-homogeneous Gaussian regression (NGR) method implemented for operational use in the Finnish Meteorological Institute (FMI). In addition to the ensemble mean and standard deviation, we also tested how the inclusion of station-specific predictors to the model affected calibration performance. Keeping operational limitations in mind, the calibration followed the operational post-processing approach. Calibration coefficients were estimated by using the past 30 day forecasts and observations pairs from all available stations, as storing the huge amount of historical data would be too demanding from the operational perspective. Furthermore, the calibration coefficients were updated once a week, and separate coefficients were estimated for each lead time.

FIGURE 8 The distribution of the continuous ranked probability skill score (CRPSS) at stations in Finland averaged over lead times from 3 to 72 hr with three different training domains and calibration models. The average CRPSS is shown above each violin plot. The reference model is the raw ensemble.

5 | CONCLUSIONS

This paper examined the calibration of the European Centre for Medium-Range Weather Forecasts ensemble forecasts (ECMWF-ENS) for 2 m temperature with lead times up to 240 hr using the non-homogeneous Gaussian regression (NGR) method implemented for operational use in the Finnish Meteorological Institute (FMI). In addition to the ensemble mean and standard deviation, we also tested how the inclusion of station-specific predictors to the model affected calibration performance. Keeping operational limitations in mind, the calibration followed the operational post-processing approach. Calibration coefficients were estimated by using the past 30 day forecasts and observations pairs from all available stations, as storing the huge amount of historical data would be too demanding from the operational perspective. Furthermore, the calibration coefficients were updated once a week, and separate coefficients were estimated for each lead time.
Verification results showed that the raw ECMWF-ENS forecasts tend to be systematically under-dispersive and biased. Compared with the raw ensemble, the operational and two alternative versions of the calibration method improved the ensemble forecasts on average over Europe. The current operational version of the calibration method, which uses elevation as the only station-specific predictor in the NGR model, provided a robust forecast performance in terms of the continuous ranked probability skill score (CRPSS) at all lead times when compared with calibration without station-specific predictors. Furthermore, using latitude and land–sea mask (lsm) as additional station-specific predictors were most beneficial at short lead times (3–72 hr), and particularly in January. Verification statistics calculated for Finnish stations using all available stations from three different training domains (Figure 1) showed that with short lead times (3–72 hr) it was most beneficial to have a smaller training domain. The differences were statistically significant at the 5% level. With longer lead times, the best calibration results were obtained by using the whole of Europe for training, except in January.

Calibration improved the forecast skill most in March, when the ECMWF-ENS forecasts showed a consistent cold bias over the whole of Europe. On the other hand, the forecast calibration actually worsened the ensemble forecasts at many stations in January 2019 (see Figure S4 in Appendix S1 in the additional supporting information). The reason for the deterioration in the calibration performance was the strong location dependence of the forecast bias. This was seen especially in northern Finland and Sweden where the ECMWF-ENS had a strong warm bias in contrast to a cold bias seen in the other parts of Europe. These kinds of large forecast errors in northern Finland are typical for the ECMWF-ENS when really cold temperatures are observed (Hyrkkänen et al., 2016). While this bias was reduced by introducing additional station-specific predictors or limiting the training domain size, in other months the effect was in some cases actually the opposite, especially at longer lead times. This indicates that no single calibration method will provide good results at all lead times. For the cal_oper method developed here, the intended use is for post-processing the ECMWF-ENS forecast at time scales from 3 to 10 days.

Even though the operationally implemented method failed to improve ensemble forecasts in Finland in January 2019, the systematic improvements in other cases showed that it is overall beneficial and even necessary to calibrate the ECMWF-ENS forecasts for operational purposes. It is also worth remembering that calibration cannot correct any forecast errors if the training data set does not contain similar cases. To improve the proposed calibration method further, predictors such as the inversion index, terrain properties or distance to the seashore could be tested because these might help to explain forecast bias during low-temperature cases. An interesting approach would be the use of spatially varying coefficients, for example, by using regression splines in the gamlss R package. Also, several lead times could be modelled together to account for correlated errors. A straightforward way to improve the current operational post-processing system would be to update the calibration coefficients every time new forecasts become available. Also, relaxing the normality assumption on forecast errors (e.g. Gebetsberger et al., 2019) could help to calibrate better the anomalously cold temperatures. These are left as subjects for future studies.

ACKNOWLEDGEMENTS
Related to the HIRLAM-C consortium, the authors thank Alex Deckmyn (Royal Meteorological Institute of Belgium) and John Bjornar Bremnes and Thomas Nipen (Norwegian Meteorological Institute) for providing the original forecast and calibration codes.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest.

ORCID
Kaisa Ylinnen https://orcid.org/0000-0002-1156-0092
Olle Räty https://orcid.org/0000-0002-6766-1167
Marko Laine https://orcid.org/0000-0002-5914-6747

REFERENCES
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.
Dabernig, M., Mayr, G.J., Messner, J.W. and Zeileis, A. (2017) Spatial ensemble post-processing with standardized anomalies. Quarterly Journal of the Royal Meteorological Society, 143, 909–916.
Diaz, M., Nicolis, O., Marin, J.C. and Baran, S. (2019) Statistical post-processing of ensemble forecasts of temperature in Santiago de Chile. Meteorological Applications, 27(1), e1818.
Feldmann, K., Scheuerer, M. and Thorarinsdottir, T.L. (2015) Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. Monthly Weather Review, 143, 955–971.
Frogner, I.-L., Andrae, U., Bojarova, J., Callado, A., Escribà, P., Feddersen, H., Hally, A., Kauhanen, J., Randriamampianina, R., Singleton, A., Smet, G., van der Veen, S. and Vignes, O. (2019) HarmonEPS—the HARMONIE ensemble prediction system. Weather and Forecasting, 34, 1909–1937.
Gebetsberger, M., Stauffer, R., Mayr, G.J. and Zeileis, A. (2019) Skewed logistic distribution for statistical temperature post-processing in mountainous areas. Advances in Statistical Climatology Meteorology and Oceanography, 5, 87–100.
Glahn, H.R. and Lowry, D.A. (1972) The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, 11, 1203–1211.

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

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.

Hagedorn, R., Hamill, T.M. and Whitaker, J.S. (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part i: two-meter temperatures. *Monthly Weather Review*, 136, 2608–2619.

Hansen, J.A. and Emanuel, K.A. (2003) Forecast 4d-var: exploiting model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 129, 1255–1267.

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.

Hyryräinen, J., Kilpinen, J., Nurmi, P., Kaurola, J. and Brockmann, M. (2016) Error characteristics of temperature forecast in Finland for the period 1979–2011 in relation to various weather patterns. *Meteorological Applications*, 23, 244–253. Available at: ehttps://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.1550.

Iversen, T., Deckmyn, A., Santos, C., Sattler, K., Bremnes, J.B., Feddersen, H. and Frognér, I.-L. (2011) Evaluation of ‘GLAMEPS’—A proposed multimodel EPS for short range forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 63, 513–530.

Lang, M.N., Lerch, S., Mayr, G.J., Simon, T., Stauffer, R. and Zelleis, A. (2020) Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Processes in Geophysics*, 27, 23–34.

Lerch, S. and Baran, S. (2017) Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66, 29–51.

Möller, A., Thorarinsdottir, T.L., Lenkoski, A. and Gneiting, T. (2015) Spatially adaptive, Bayesian estimation for probabilistic temperature forecasts. arXiv:1507.05066. 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. (2000) Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, 63, 71–116.

Pinson, P. and Messner, J.W. (2018) Application of postprocessing for renewable energy. In: *Statistical Postprocessing of Ensemble Forecasts*. Amsterdam: Elsevier, pp. 241–266.

R Core Team. (2019) R: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing https://www.R-project.org/.

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

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 507–554.

Rosgaard, M.H., Nielsen, H.A., Nielsen, T.S. and Hahmann, A.N. (2016) Probing NWP model deficiencies by statistical post-processing. *Quarterly Journal of the Royal Meteorological Society*, 142, 1017–1028.

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

Thorarinsdottir, T.L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173, 371–388.

Van Wagner, C. (1987) Development and structure of the Canadian forest fireweather index system. In Canadian Forestry Service, Forest Technical Report. Citeseer.

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*, Vol. 100 Amsterdam: Academic Press.

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

Willet, M.R., Bechtold, P., Williamson, D., Petch, J., Milton, S. and Woolnough, S. (2008) Modelling suppressed and active convection: comparisons between three global atmospheric models. *Quarterly Journal of the Royal Meteorological Society*, 134, 1881–1896.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Ylinen K, Räty O, Laine M. Operational statistical postprocessing of temperature ensemble forecasts with station-specific predictors. *Meteorol Appl.* 2020;27:e1971. https://doi.org/10.1002/met.1971