Received: 14 September 2018 | Revised: 17 May 2019 | Accepted: 19 June 2019 | Published on: 17 July 2019

DOI: 10.1002/met.1818

RESEARCH ARTICLE

Meteorological Applications RMetS

Statistical post-processing of ensemble forecasts of temperature in Santiago de Chile

Mailiu Díaz[1] | **Orietta Nicolis**[1,2] | **Julio César Marín**[3,4] | **Sándor Baran**[5]

[1]Department of Statistics, University of Valparaíso, Valparaíso, Chile

[2]Department of Engineering Science, Faculty of Engineering, Andres Bello University, Viña del Mar, Chile

[3]Department of Meteorology, University of Valparaíso, Valparaíso, Chile

[4]Interdisciplinary Center for Atmospheric and Astro-statistical Studies (CEAAS), University of Valparaíso, Valparaíso, Chile

[5]Department of Applied Mathematics and Probability Theory, Faculty of Informatics, University of Debrecen, Debrecen, Hungary

Correspondence
Sándor Baran, Department of Applied Mathematics and Probability Theory, Faculty of Informatics, University of Debrecen, Kassai út 26, H-4028, Debrecen, Hungary.
Email: baran.sandor@inf.unideb.hu

Funding information
Comisión Nacional de Investigación Científica y Tecnológica, Grant/Award Number: 21150227; Bolyai János Research Scholarship of the Hungarian Academy of Sciences; National Laboratory for High Performing Computer (NLHPC), Grant/Award Number: ECM-02; Interdisciplinary Center of Atmospheric and Astro-Statistical Studies; National Research, Development and Innovation Office, Grant/Award Number: NN125679

Abstract

Modelling forecast uncertainty is a difficult task in any forecasting problem. In weather forecasting a possible solution is the use of forecast ensembles, which are obtained from multiple runs of numerical weather prediction models with various initial conditions and model parametrizations to provide information about the expected uncertainty. Currently all major meteorological centres issue forecasts using their operational ensemble prediction systems. However, it is a general problem that the spread of the ensemble is too small compared to observations at specific sites resulting in under-dispersive forecasts, leading to a lack of calibration. In order to correct this problem, various statistical calibration techniques have been developed in the last two decades. In the present work different post-processing techniques were tested for calibrating nine member ensemble forecasts of temperature for Santiago de Chile, obtained by the Weather Research and Forecasting model using different planetary boundary layer and land surface model parametrizations. In particular, the ensemble model output statistics and Bayesian model averaging techniques were implemented and, since the observations are characterized by large altitude differences, the estimation of model parameters was adapted to the actual conditions at hand. Compared to the raw ensemble, all tested post-processing approaches significantly improve the calibration of probabilistic forecasts and the accuracy of point forecasts. The ensemble model output statistics method using parameter estimation based on expert clustering of stations (according to their altitudes) shows the best forecast skill.

KEYWORDS

Bayesian model averaging, ensemble model output statistics, ensemble post-processing, probabilistic forecasting, temperature forecast

1 | INTRODUCTION

The central zone of Chile, located between 32 ° S and 37 ° S latitude, has a semi-arid Mediterranean climate as a result of the influence of topographic barriers and the southeast Pacific anticyclone located over the cool Pacific Ocean, which generates a persistent inversion layer in the lowest few hundred metres of the atmosphere (Burger *et al.*, 2018). In particular, Santiago de Chile surrounded by high mountains has its own special micro climate making accurate weather forecasts even more complicated.

Obtaining reliable forecasts of surface temperature has a large impact in many fields such as renewable energy, air quality and radiative transfer. These forecasts are typically produced with the help of numerical weather prediction models, which provide predictions at high spatial and temporal resolutions. In the last few years in Chile, serious efforts have been made in evaluating meteorological models to improve temperature forecasts, among other variables (Cortés and Curé, 2011; Saide *et al*., 2011; Pozo *et al*., 2016; González and Garreaud, 2019). However, the outputs of these numerical weather prediction models are subject to an intrinsic uncertainty, which for a specific region can be quantified by running the models with different initial conditions and parametrizations opening the door for ensemble forecasting (Bauer *et al*., 2015). Currently, all major meteorological centres generate ensemble forecasts using their operational ensemble prediction systems (EPSs). Examples are the 51 member EPSs of the European Centre for Medium-Range Weather Forecasts (Molteni *et al*., 1996; Leutbecher and Palmer, 2008) and the 30 member Consortium for Small-scale Modelling EPSs of the German Meteorological Service (Gebhardt *et al*., 2011). A general problem with many operational EPSs is that the spread of the ensemble is too small, and the observations at specific sites often fall outside the range of ensemble members. Such an under-dispersive forecast characteristic results in a lack of calibration (see foe example Buizza *et al*., 2005; Vannitsem *et al*., 2018).

A possible way of improving ensemble member forecasts is the use of some form of statistical post-processing. In the last 15 years several different statistical calibration techniques have been developed (for an overview see Williams *et al*., 2014; Vannitsem *et al*., 2018) including non-homogeneous regression or ensemble model output statistics (EMOS) (Gneiting *et al*., 2005) and ensemble BMA (Raftery *et al*., 2005), a post-processing approach based on the idea of Bayesian model averaging. Both methods provide full predictive distribution of the weather variable at hand. The EMOS predictive distribution is specified by a parametric distribution with parameters connected to the ensemble members *via* appropriate link functions, whereas BMA applies mixture distributions with components corresponding to the ensemble members. Given a predictive distribution, either its mean or its median can be considered as point forecasts, and probabilities of various events can also be easily calculated. EMOS and BMA models corresponding to various weather quantities differ in the parametric laws on which they are based. Temperature and pressure forecasts are fitted well by a normal distribution (Gneiting *et al*., 2005; Raftery *et al*., 2005), wind speed requires a non-negative and skewed distribution such as truncated normal (Thorarinsdottir and Gneiting, 2010; Baran, 2014), log-normal (Baran and Lerch, 2015) or gamma (Sloughter *et al*., 2010), whereas to calibrate accumulated precipitation a discrete–continuous model with point mass at zero is required (Sloughter *et al*., 2007; Scheuerer, 2014; Scheuerer and Hamill, 2015; Baran and Nemoda, 2016).

In the present work various post-processing models for calibrating ensemble forecasts of temperature in Santiago city were evaluated. The ensemble members correspond to nine Weather Research and Forecasting (WRF) (Powers *et al*., 2017) model configurations with three nested domains. According to the best knowledge of the authors, no studies have been published yet with the aim of improving the quality of surface temperature predictions in Chile by statistical calibration based on ensemble post-processing techniques. Moreover, as the results show, the location of the ensemble domain with large altitude differences requires some adaptation of the post-processing approach to the actual conditions at hand.

The paper is organized as follows. A description of WRF configurations, data from meteorological stations and their preliminary statistical analysis is provided in Section 2. Section 3 describes the post-processing models and applied methods of model verification, whereas the results of the statistical post-processing are given in Section 4. Finally, Section 5 concludes the paper with a summary of the major findings and a discussion of possible future areas of research.

2 | WRF CONFIGURATIONS AND DATA DESCRIPTION

The WRF model was employed to generate nine different simulations resulting in a nine member forecast ensemble for surface temperature (K) for the period between October 1, 2017, and January 30, 2018. WRF model outputs during the study period were generated at 3 hr intervals. The corresponding verifying observations were obtained from 19 meteorological stations around Santiago city.

2.1 | WRF simulations

Three model nested domains (see Figure 1a), at 18, 6 and 2 km horizontal resolutions, were employed in the simulations using version 3.7.1 of the Advanced Research WRF core (ARW-WRF) (Skamarock *et al*., 2008). Results from the highest resolution domain (d3) were used in this study with a superficial area of 208 km × 208 km.

Data from the global final analysis (FNL) (RDA CISL ds083.3; NCEP, 2015) were used to create the initial state and to update the boundary conditions for the regional WRF simulations. The FNL at 0.25° by 0.25° horizontal resolution is available every 6 hr at 0000, 0600, 1200 and 1800 UTC

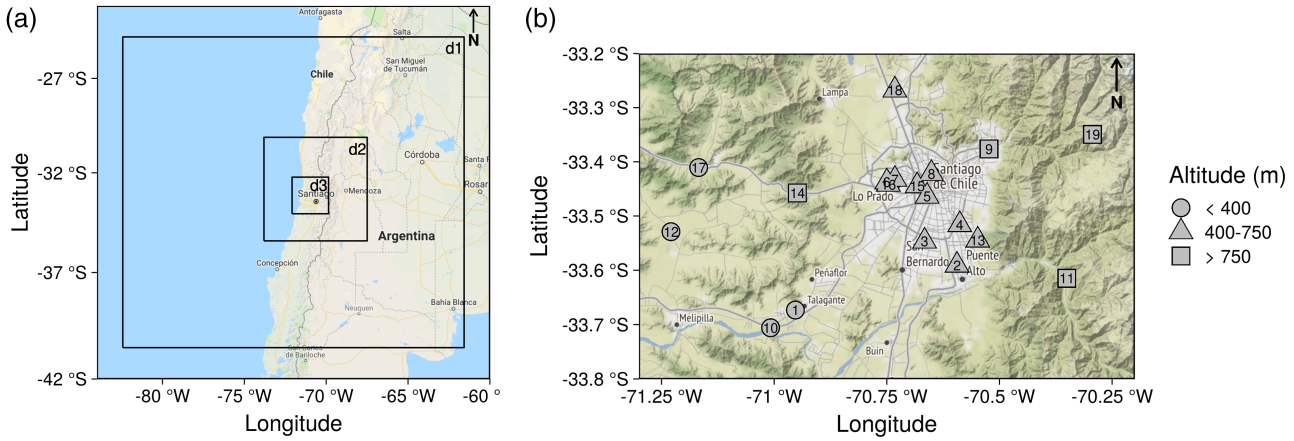

FIGURE 1 (a) Weather Research and Forecasting domain configuration domains d1, d2 and d3; (b) domain d3 with the superposed terrain elevation (m) and the location of the 19 meteorological stations used in the study

TABLE 1 Description of the parametrization set used on each of the nine ensemble members

Member	LSM	Surface	PBL	Cumulus	Microph.	LW rad	SW rad
1	Noah	MYNN	MYNN	Kain-F	WSM3	RRTMG	RRTMG
2	Noah	MYJ	MYJ	Kain-F	WSM3	RRTMG	RRTMG
3	Noah	YSU	YSU	Kain-F	WSM3	RRTMG	RRTMG
4	Pleim-Xiu	MYJ	MYJ	Kain-F	WSM3	RRTMG	RRTMG
5	Pleim-Xiu	YSU	YSU	Kain-F	WSM3	RRTMG	RRTMG
6	5-layer	MYJ	MYJ	Kain-F	WSM3	RRTMG	RRTMG
7	5-layer	YSU	YSU	Kain-F	WSM3	RRTMG	RRTMG
8	5-layer	MYNN	MYNN	Kain-F	WSM3	RRTMG	RRTMG
9	Pleim-Xiu	MYNN	MYNN	Kain-F	WSM3	RRTMG	RRTMG

Cumulus, the convective scheme; Kain-F, Kain–Fritsch; LSM, land surface model; LW Rad and SW Rad, long wave and short wave radiation schemes; Microph., microphysics scheme; MYJ, Mellor–Yamada–Janjic; MYNN, Mellor–Yamada Nakanishi and Ninno; PBL, planetary boundary layer scheme; RRTMG, rapid radiative transfer model; WSM3, Weather Research and Forecasting single-moment 3-class; YSU, Yonsei University.

where the 0000 UTC analysis is used to initialize the WRF model runs. An analysis nudging was implemented in the outer domain (d1) to provide better boundary conditions during the simulation period. This setting cannot be used operationally but just as a research case study. However, in the future, global forecast system (GFS) forecasts with a 3 hr temporal resolution could be used to replace FNL to run an operational GFS-WRF setting with 3 hr nudging. The simulations included 44 vertical levels with variable resolution between 60 and 200 m and eight levels within the first kilometre in the vertical. The validation of raw and post-processed ensemble forecasts was performed with the 24 hr forecasts of each day (from 0000 to 2100 UTC). Table 1 describes the parametrizations used in each simulation.

The members differ from each other in the applied planetary boundary layer (PBL) and land surface model parametrizations. The Mellor–Yamada–Janjic (Janjić, 1994), Yonsei University (Hong *et al*., 2006) and Mellor–Yamada Nakanishi and Ninno 2.5 level (Nakanishi and Niino, 2006) schemes are used to represent the PBL and surface layer processes. The land surface processes are represented by the five-layer (Dudhia, 1996), Noah (Chen and Dudhia, 2001) and Pleim–Xiu (Pleim and Xiu, 2003) schemes. The rest of the parametrizations are kept the same in all simulations. The Kain–Fritsch (Kain, 2004) cumulus parametrization is used to represent the convective processes in domain 1 (18 km) and the WRF single-moment 3-class (Hong *et al*., 2004) scheme is used to represent microphysics processes. Finally, to represent the long wave and short wave radiative processes the rapid radiative transfer model (Iacono *et al*., 2005) is applied.

2.2 | Station observations

Hourly mean surface temperature observations (K) every 3 hr for the period between October 1, 2017, and January 30, 2018, were obtained for 19 meteorological stations whose descriptions are shown in Table 2 (see also

TABLE 2 Geographical co-ordinates in decimal degrees and altitude (m) of monitoring stations and altitude of the nearest grid-point to the station in the Weather Research and Forecasting (WRF) simulation

No.	Station	Latitude (°S)	Longitude (°W)	Altitude (m)	WRF height (m)
1	Talagante	33.67	70.95	390	281.97
2	Puente Alto	33.59	70.59	670	667.06
3	El Bosque	33.55	70.67	580	559.47
4	La Florida	33.52	70.59	601	590.45
5	Parque O'Higgins	33.46	70.66	549	506.83
6	Pudahuel	33.44	70.75	494	462.78
7	Cerro Navia	33.43	70.73	500	466.98
8	Independencia	33.42	70.65	560	548.83
9	Las Condes	33.38	70.52	798	785.95
10	El Paico	33.71	71.01	275	222.30
11	San José Guayacán	33.61	70.35	928	1,297.73
12	Chorombo hacienda	33.53	71.23	145	126.26
13	Aguas Andinas	33.54	70.55	665	758.60
14	Lo Prado	33.46	70.95	1,068	812.29
15	Quinta Normal	33.44	70.68	534	496.49
16	San Pablo	33.44	70.75	490	461.43
17	Curacaví	33.41	71.17	208	237.14
18	Lo pinto	33.27	70.73	512	483.92
19	El Colorado	33.35	70.29	2,750	2,940.90

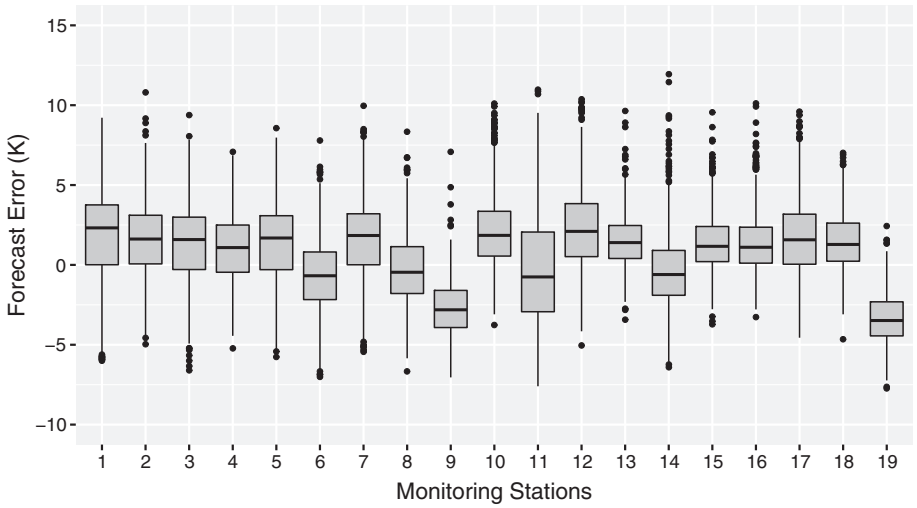

FIGURE 2 Median forecast error of the ensemble at the monitoring stations listed in Table 2

Figure 1b). The stations are managed by the Dirección Meteorológica de Chile, which is the government agency responsible for managing the meteorological data (available online at http://www.meteochile.gob) and provides operational weather forecasts in the country, and by the National System for Air Quality (Environmental Ministry, see https://sinca.mma.gob.cl/). It should be noted that the three stations Las Condes (9), La Florida (4) and Chorombo Hacienda (12) have missing values in more than 25% of their data.

2.3 | Predictive performance of the raw ensemble

To verify the predictive performance of the raw WRF ensemble, temperature forecasts from the nearest grid-point to the location of each weather station were extracted from the WRF domain d3 and compared to the corresponding observations. As Table 2 and Figure 1b show, the topography within the domain used in this study is very complex

causing difficulties in obtaining reliable forecasts. Figure 2 provides box-plots of the median forecast error of the ensemble for each station. The raw ensemble systematically underestimates temperature at stations located at higher altitudes (stations 9, 11, 14 and 19). This should obviously be taken into account during the calibration process. A number of studies have found cold biases in near-surface temperature forecasts from the WRF model in different mountainous regions in Chile and other parts of the world using similar options for PBL and land surface model schemes as those used in this study (Ruiz *et al*., 2010; Marín *et al*., 2013; Massey *et al*., 2016; González and Garreaud, 2019). The temperature underestimation may be the result of misrepresentations in the real orography or the near-surface moisture in the model (Massey *et al*., 2016). Table 2 indicates that the largest differences between the real and model topography are shown for stations 11 (360 m), 14 (256 m) and 19 (190 m). These differences can only explain parts of the large mean bias shown for station 19. Other factors, e.g. land-use misinterpretation, local weather phenomena but also the choice of the station clusters, may contribute to the bias, as demonstrated for this station as well as for stations 9 and 11, but it has not been further investigated.

As shown in Figure 3a providing the box-plots of forecast errors of the individual ensemble members over all available forecast cases, there is also some variation in the performance of the different ensemble members. Ensemble members 7 and 8 underperform the other seven ensemble members as they tend to have larger forecast errors. Furthermore, the forecast error varies with forecast time, as Figure 3b indicates. The box-plots of the diurnal evolution of the ensemble mean forecast errors show that simulations mainly overestimate the observed temperatures except at 1200 UTC, where they slightly underestimate it, and at 1500 UTC, where the forecasts seem to be unbiased. The dependence of the bias on the validation time might be related to the misrepresentation of the real orography in the model. The diurnal temperature variation is strongly influenced by solar radiation, which strongly varies in sites located in complex terrain due to topographic shading and variations in slope orientations (Zhang *et al*., 2018). Therefore, the time of sunrise in those stations might be misrepresented, causing larger errors at the above mentioned hours.

In order to get a first overview of the raw ensemble calibration, the verification rank histogram can be examined (Figure 4), which represents the histogram of ranks of validating observations with respect to the corresponding ensemble forecasts computed for all forecast cases (see for example Wilks, 2011, section 7.7.2). For properly calibrated ensemble forecasts, the ranks should be uniformly distributed, which is clearly not the case in Figure 4, calling for some form of statistical post-processing.

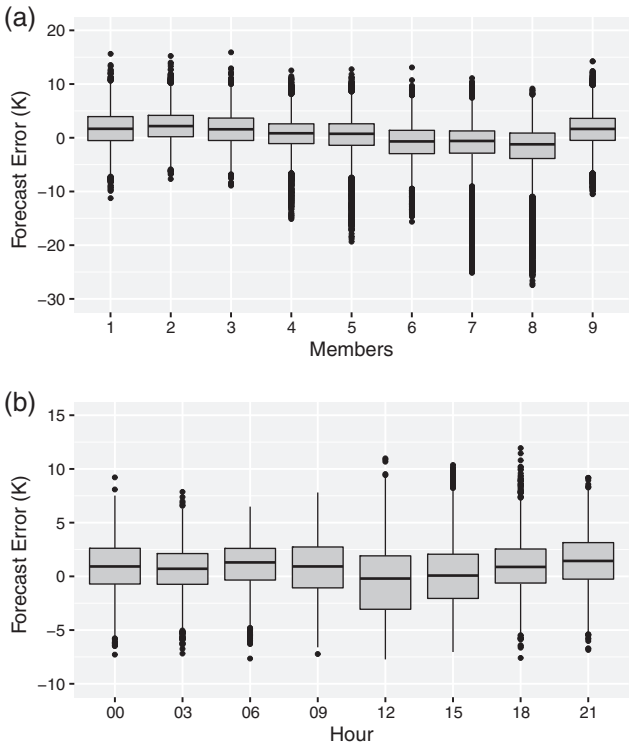

FIGURE 3 Forecast error (a) of the different ensemble members for all dates and time points; (b) of the ensemble mean for different time points for all dates

3 | MODELS

As mentioned in Section 1, a normal distribution is often suitable for temperature modelling. Hence, the normal

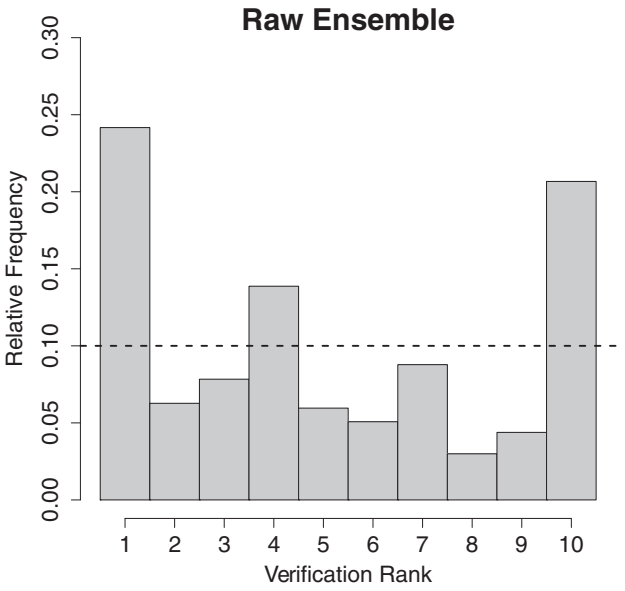

FIGURE 4 Verification rank histogram of raw ensemble forecasts for the period October 1, 2017–January 30, 2018

EMOS and BMA models suggested by Gneiting *et al.* (2005) and Raftery *et al.* (2005), respectively, were used to calibrate the ensemble temperature forecasts from the WRF model. In what follows, let $f_1, \ldots, f_9$ denote the forecasts of the nine ensemble members for a given location and time point.

3.1 | EMOS model for temperature

The normal EMOS predictive distribution for the nine WRF ensemble members has the form:

$$\mathcal{N}\left(a_0 + a_1 f_1 + \cdots + a_9 f_9, b_0 + b_1 S^2\right) \tag{1}$$

where

$$S^2 := \frac{1}{8}\sum_{k=1}^{9}\left(f_k - \bar{f}\right)^2$$

is the ensemble variance, with $\bar{f}$ denoting the ensemble mean. Location parameters $a_0 \in \mathbb{R}$ and $a_1, \ldots, a_9 \geq 0$ and scale parameters $b_0, b_1 \geq 0$ were estimated by optimizing the value of a proper scoring rule (see Section 3.3) over the training data consisting of ensemble forecasts and verifying observations from the preceding n days.

3.2 | BMA model for temperature

EMOS calibration is efficient in the case of unimodal distributions. However, it cannot provide an appropriate model for weather variables following distributions with several modes. In such situations BMA modelling using mixture distributions might outperform EMOS models. The normal BMA model of Raftery *et al.* (2005) for calibrating temperature forecasts in this case results in a predictive probability density function:

$$p(x|f_1,\ldots,f_9) := \sum_{k=1}^{9}\omega_k \frac{1}{\sigma}\varphi\left(\frac{x-\beta_{0,k}-\beta_{1,k}f_k}{\sigma}\right) \tag{2}$$

where φ denotes the probability density function of the standard normal distribution. Here each weight ω_k considers the relative performance of the corresponding kth ensemble member f_k during the training period, and the weights have to fulfil the condition $\sum_{k=1}^{9}\omega_k = 1$, $\omega_k \geq 0$. The linear form $\beta_{0,k} + \beta_{1,k}f_k$ of the average over each component probability density function is responsible for the bias correction; however, cases $\beta_{1,k} = 1$ (additive bias correction) and $\beta_{0,k} = 0$, $\beta_{1,k} = 1$ (no bias correction) might also be considered. The last two approaches might be helpful in situations where the raw ensemble is unbiased and additional bias correction might introduce unnecessary extra errors (see for example Baran *et al.*, 2014).

Similar to the EMOS approach, location parameters $\beta_{0,k}$, $\beta_{0,1}$, weights ω_k, $k = 1, \ldots, 9$, and scale σ were estimated using appropriate training data. Location parameters were obtained by regressing the validating observations on the ensemble members, whereas weights and scale were estimated using a maximum likelihood approach where the likelihood function was maximized with the help of the expectation maximization algorithm for mixtures (see for example McLachlan and Krishnan, 1997). Note that BMA modelling was performed with the help of the ensembleBMA R package (Fraley *et al.*, 2011), whereas EMOS models were fitted using custom codes tailored to the problems at hand.

3.3 | Verification scores

The aim of statistical post-processing is to maximize the sharpness of the predictive distribution subject to calibration (Gneiting *et al.*, 2007), where the former refers to the concentration of the forecast distribution and the latter to the consistency between predicted probabilities and observed relative frequencies. These goals can be assessed simultaneously with the help of scoring rules (Gneiting and Raftery, 2007) assigning numerical values to pairs of forecast distributions and validating observations. One of the most widely used proper scoring rules in atmospheric sciences is the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007). For a predictive cumulative distribution function (CDF) $F(y)$ and observation x the CRPS is defined as:

$$\mathrm{CRPS}(F,x) := \int_{-\infty}^{\infty}\left\{F(y) - 1_{\{y \geq x\}}\right\}^2 \mathrm{d}y = E|X-x| - \frac{1}{2}E|X-X'|$$

where 1_H denotes the indicator of a set H, while X and X' are independent random variables with CDF F and finite first moment. Note that the CRPS can be expressed in the same units as the observation and is a negatively oriented scoring rule, the smaller the better.

Furthermore, the calibration of a predictive distribution can be investigated using the coverage of the $(1-\alpha)100\%$, $\alpha \in (0,1)$, central prediction interval. The coverage is defined as the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive CDF and level α should be chosen to match the nominal coverage of the raw ensemble, which is 80% for the nine member ensemble at hand. As the coverage of a calibrated predictive distribution should be around $(1-\alpha)100\%$, the suggested choice of α allows direct comparisons with the raw ensemble.

The improvement in calibration with respect to the raw ensemble can also be demonstrated with the help of probability integral transform (PIT) histograms (Wilks, 2011). The PIT is the value of the predictive CDF evaluated at the validating observation and in the case of proper calibration it should follow a uniform law on the [0,1] interval. Hence, the PIT histogram is the continuous counterpart of the verification rank histogram of the raw ensemble.

Finally, the predictive performance of point forecasts such as the ensemble median and mean was evaluated with the help of the mean absolute error (MAE) and the root mean squared error (RMSE). Note that the former is optimal for the median, whereas the latter is optimal for the mean (Gneiting, 2011; Pinson and Hagedorn, 2012).

3.4 | Training data

The choice of appropriate training data is essential for estimating the parameters of the EMOS and BMA models given by Equations 1 and 2. In many situations, rolling training periods are applied and there are two main approaches to choosing forecast cases for training (Thorarinsdottir and Gneiting, 2010). In the local approach, parameters for a given station are estimated only from the data of that particular station. This approach usually results in a very good model fit, provided the training period is long enough to avoid numerical issues in parameter estimation (see for example Hemri *et al.*, 2014). In contrast, regional EMOS and BMA estimate parameters using all available forecast cases from the training period; thus, all stations in the forecast domain share the same set of parameters. In this way shorter training periods can be used, but the regional approach is not suitable for large heterogeneous domains. Recently, Lerch and Baran (2017) proposed a third, semi-local approach, which combines the advantages of local and regional forecasting. Training data for a given station are augmented with data for stations with similar characteristics, e.g. by clustering the stations using feature vectors determined by station climatology and/or ensemble forecast errors during the training period. Within a given cluster a regional parameter estimation is performed, but note that clusters may vary as the training window slides (for more details see Lerch and Baran, 2017). Finally, an expert clustering of the monitoring stations can also be performed based on their location or other covariates and the parameters within clusters can be estimated regionally.

4 | RESULTS

In order to exclude the natural daily variation in temperature, the EMOS and BMA calibrations of the WRF ensemble forecasts described in Section 2.1 were performed separately for each forecast validation time point. Since the ensemble is composed of nine members, EMOS post-processing requires a total of 12 parameters to be estimated, whereas for BMA the number of free parameters is 27. The dataset at hand covers only 122 calendar days which for both approaches makes the local estimation of parameters impossible. Regional estimation also requires at least a 6 day training period (6 days × 19 stations = 114 forecast cases) for EMOS and at least a 15 day training period (285 forecast cases) for BMA in order to have about 10 times more forecast cases than parameters.

The length of the optimal training period is determined by calibrating the ensemble forecasts using training periods of length 10, 15, …, 60 days and comparing the predictive performance on the verification period November 30, 2017–January 30, 2018 (62 calendar days). Figure 5 shows the mean CRPS values of EMOS and BMA predictive distributions as functions of the training period length for all time points. Both models have the best predictive performance at time points 0300 and 0600 UTC, providing slightly higher CRPS values for 0000, 0900, 1800 and 2100 UTC, whereas the worst forecast skill corresponds to 1200 and 1500 UTC. However, these results are partially in line with Figure 3b and might be explained by differences in the accuracy of ensemble forecasts for different periods of the day. Note that the curves present their minima at day 10 and, except for the curves corresponding to 1200 and 1500 UTC, they do not show much variability. The MAE and RMSE values of EMOS and BMA median and mean forecasts (not reported), respectively, are very consistent with the CRPS and do not change the overall picture. In many cases short training periods are preferred, but also the minimum number of forecast cases should be kept in mind for parameter estimation and a sufficiently large training period should be selected, which depends on the dataset and the number of stations. For this study, a training period of 20 days was chosen for

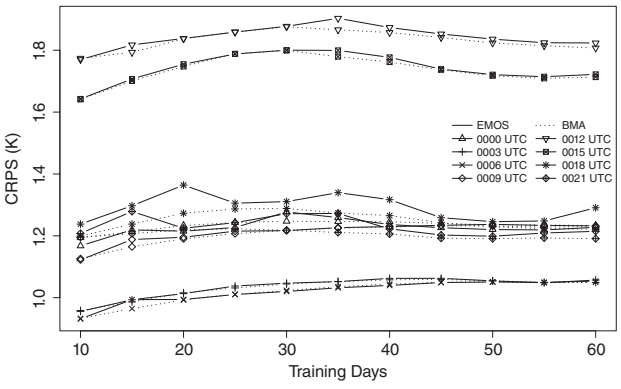

FIGURE 5 Mean continuous ranked probability score (CRPS) of ensemble model output statistics (EMOS) and Bayesian model averaging (BMA) predictive distributions for the period November 30, 2017–January 30, 2018

calibrating the WRF forecasts (see Figure 5) as it might also be appropriate for semi-local estimation of EMOS parameters using two or three clusters of observation stations. In this way the predictive performance of EMOS and BMA post-processed forecasts can be tested on temperature data from the period between October 21, 2017, and January 30, 2018 (102 calendar days).

Besides regional estimation of BMA and EMOS parameters, clustering based semi-local EMOS parameter estimation was also tested with two and three clusters and 24 features. Half of the features are obtained as equidistant quantiles of the climatological CDF and the other half as equidistant quantiles of the empirical CDF of the forecast error of the ensemble mean over the training period (Lerch and Baran, 2017). However, this approach often produces very unbalanced cluster sizes even for two clusters. In almost 28% of the cases, station 19 alone forms a separate cluster and the other cluster consists of the remaining stations. In particular, for 1800 UTC this was the case on 76 out of 102 days in the training period. This uneven clustering is in line with the bad ensemble forecast skill at station 19 (Figure 2). Having a single station in a cluster means local parameter estimation for that particular station, resulting in numerical issues in the optimization procedure.

Instead of grouping the stations dynamically based on feature vectors, some form of expert clustering might be tried and, for the ensemble domain at hand (see Figure 1b), altitude might be a reasonable covariate. The chosen altitude regions resulting in three clusters are as follows: under 400 m (stations 1, 10, 12, 17); between 400 and 750 m (stations 2, 3, 4, 5, 6, 7, 8, 13, 15, 16, 18); above 750 m (stations 9, 11, 14, 19). Note that, at stations in the third cluster, WRF forecasts systematically underestimate temperature (see Figure 2 and the corresponding discussion in Section 2.3).

In this way, for post-processing ensemble forecasts, BMA and EMOS models with regional parameter estimation and the EMOS model with expert clustering (EMOS-C) were considered using a 20 day rolling training period. Table 3 shows the mean CRPS of probabilistic forecasts, the RMSE of mean and the MAE of median forecasts and the coverage of 80% central prediction intervals separately for the different time points and for the overall verification period. Note that all post-processing approaches outperform the raw ensemble in terms of all scores for all forecast validation time points except for the coverage at 1800 UTC; however, even for that time point, the values are not far from the nominal 80%. From the three competing post-processing approaches, EMOS-C shows the best predictive performance followed by BMA and EMOS. However, the differences between the scores of the three post-processing approaches are minor.

TABLE 3 Mean CRPS of probabilistic forecasts and RMSE of mean and MAE of median forecasts, and coverage of 80% central prediction intervals

		Forecast validation time (UTC)								
Scores	**Models**	**0000**	**0300**	**0600**	**0900**	**1200**	**1500**	**1800**	**2100**	**Overall**
CRPS (K)	Ensemble	1.728	1.570	1.743	1.773	2.135	1.839	1.604	1.764	1.769
	EMOS	1.186	1.012	1.017	1.219	1.775	1.672	1.309	1.260	1.306
	EMOS-C	1.105	0.955	0.992	1.218	1.686	1.505	1.163	1.180	1.225
	BMA	1.197	1.018	1.019	1.221	1.778	1.668	1.247	1.254	1.299
RMSE (K)	Ensemble	2.794	2.569	2.817	2.835	3.265	3.297	3.041	3.019	2.963
	EMOS	2.092	1.791	1.808	2.156	3.121	2.988	2.418	2.228	2.370
	EMOS-C	1.951	1.692	1.755	2.149	2.996	2.697	2.111	2.114	2.222
	BMA	2.097	1.795	1.810	2.156	3.118	2.959	2.236	2.212	2.342
MAE (K)	Ensemble	2.255	1.903	2.153	2.312	2.712	2.496	2.244	2.500	2.321
	EMOS	1.683	1.417	1.437	1.729	2.557	2.367	1.828	1.788	1.850
	EMOS-C	1.571	1.340	1.393	1.727	2.415	2.104	1.592	1.659	1.725
	BMA	1.693	1.420	1.443	1.732	2.533	2.357	1.745	1.778	1.837
Cover (%)	Ensemble	51.15	58.14	57.05	61.14	40.46	69.04	80.98	67.18	60.63
	EMOS	75.63	75.56	76.36	76.66	77.30	77.59	75.37	78.36	76.60
	EMOS-C	76.40	74.25	75.53	75.29	76.86	73.84	72.62	76.88	75.21
	BMA	74.42	76.33	77.29	75.12	74.49	75.61	76.09	76.44	75.72

BMA, Bayesian model averaging; CRPS, continuous ranked probability score; EMOS, ensemble model output statistics; EMOS-C, EMOS with clustering; MAE, mean absolute error; RMSE, root mean squared error.

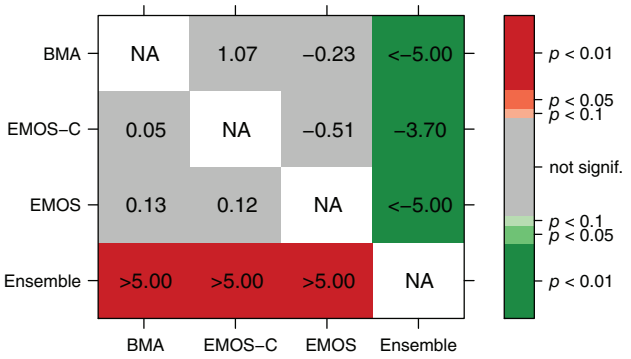

FIGURE 6 Values of the test statistics of the Diebold–Mariano test for equal predictive performance based on the continuous ranked probability score (upper triangle) and absolute error of the median forecast (lower triangle). Negative/positive values indicate a superior predictive performance of the forecast given in the row/column label, whereas the green/red background indicates significant differences

The statistical significance of differences between overall mean CRPS and MAE values was investigated with the help of the two-sided Diebold–Mariano (Diebold and Mariano, 1995) test of equal predictive performance, as this test takes into account temporal dependence (for details of the application see for example Baran and Lerch, 2018). Figure 6 shows the values of the Diebold–Mariano test statistics based on the CRPS and the absolute error of the median for all pairwise comparisons of forecasts. All calibration methods result in significant improvement compared to the raw ensemble; however, the forecast skill of the three post-processing approaches does not differ significantly. This means that, for the WRF forecasts at hand, the use of the more complex BMA model for calibration with many parameters does not necessarily pay off. Use of the simple EMOS approach should be considered instead, possibly grouping the stations in a careful way.

The same conclusions can be drawn from the overall PIT histograms plotted in Figure 7. The improvement in calibration compared to the raw ensemble is obvious; however, all three PIT histograms are slightly U-shaped indicating a very small under-dispersion (see also the overall coverage values of Table 3). Unfortunately, the Kolmogorov–Smirnov test rejects the uniformity of the PIT values at a 5% level in all three cases; however, the mean p values of 1,000 samples from PIT, each of size 1,000, given in Table 4, nicely reflect the shapes of the PIT histograms of Figure 7.

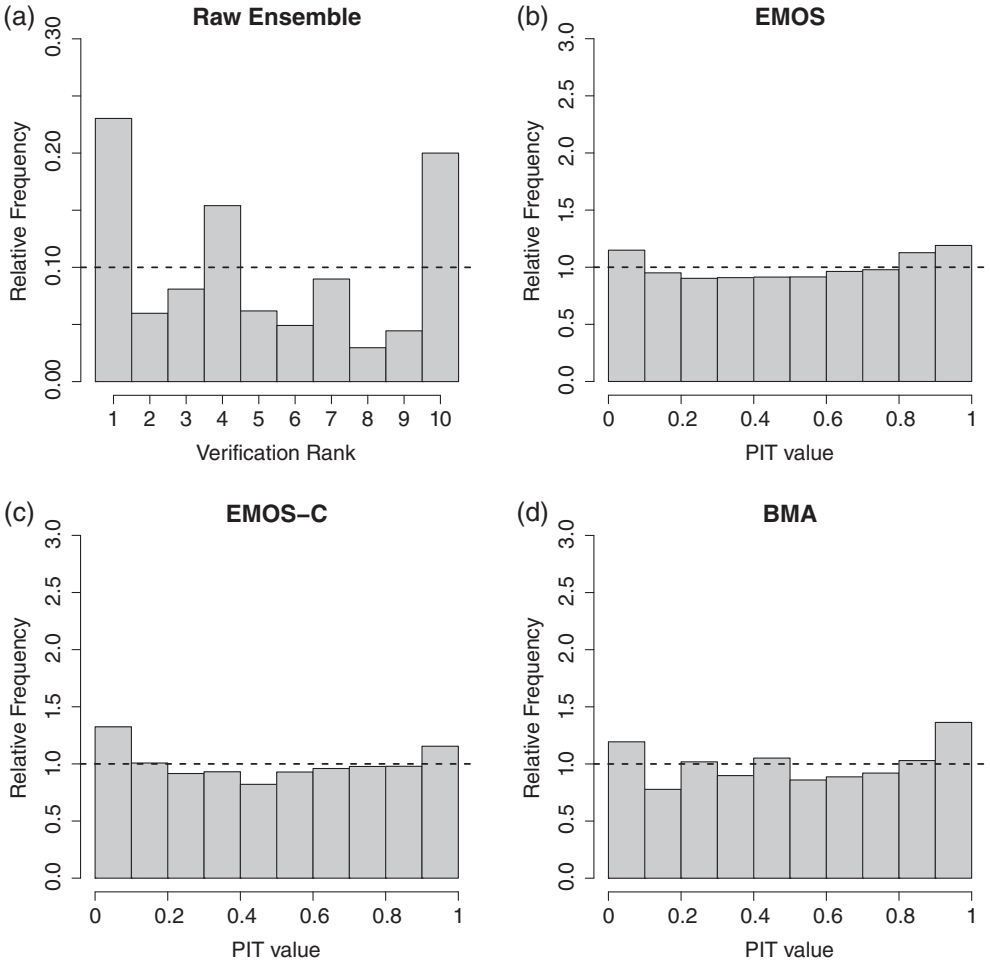

FIGURE 7 (a) Verification rank histogram of the raw ensemble, and probability integral transform (PIT) histograms of post-processed forecasts for the period October 21, 2017–January 30, 2018: (b) ensemble model output statistics (EMOS), (c) EMOS with clustering (EMOS-C), (d) Bayesian model averaging (BMA)

TABLE 4 *p* values of Kolmogorov–Smirnov tests for uniformity of PIT values

Model	EMOS	EMOS-C	BMA
Mean *p* value	0.119	0.104	0.058

Average of 1,000 random samples of size 1,000 each.
BMA, Bayesian model averaging; EMOS, ensemble model output statistics; EMOS-C, EMOS with clustering; PIT, probability integral transform.

5 | CONCLUSIONS

Two types of statistical post-processing methods are applied to 3 hr ensemble forecasts of near surface temperature for Santiago de Chile produced by separate runs of the Weather Research and Forecasting (WRF) model with nine different configurations. One day ahead predictions for different forecast validation time points are treated separately. The predictive performance of the ensemble model output statistics (EMOS) and Bayesian model averaging (BMA) models with regional parameter estimation using a 20 day training period is investigated. This optimal length of the training period is a result of a detailed data analysis. Besides the regional models, the forecast skill of an EMOS approach with parameter estimation based on expert clustering of observation stations using their altitude data, referred to as EMOS-C, is also tested. Compared to the raw ensemble, all post-processing methods for all forecast validation time points result in a significant decrease in continuous ranked probability score (CRPS) values of probabilistic forecasts and the mean absolute error (MAE) and root mean squared error (RMSE) values of point forecasts. In addition, they also yield a substantial improvement in calibration. From the competing calibration approaches, EMOS-C produces the smallest score values for all forecast validation time points; however, the differences in overall mean CRPS and MAE of post-processing models are not significant. Thus, it can be concluded that post-processing of WRF ensemble forecasts for temperature significantly improves the calibration of probabilistic forecasts and accuracy of point forecasts. However, there is still space for further improvements, e.g. by considering models including spatial dependence *via* state of the art approaches such as the Markovian EMOS (Möller *et al*., 2015), ensemble copula coupling (Schefzik, 2016a, 2016b) and the spatial extensions of BMA and EMOS suggested by Feldmann *et al*. (2015), which appear to be very suitable for the dataset at hand.

Currently, a number of meteorological centres run an ensemble of global simulations several times per day to produce global weather forecasts. For example, an ensemble of Global Forecast System forecasts is available at $0.25° \times 0.25°$ or 27×27 km horizontal resolution (at the Equator). The WRF configuration used in this study generates weather forecasts at 2×2 km, which should improve the near-surface forecasts in the region from a global model since they cover different topographic and land-use features that should be better represented in the high-resolution WRF simulations. However, it would be very interesting to apply post-processing methods such as those used in this study to a number of Global Forecast System ensemble members to determine whether their performance strongly differs from those obtained in this study. That would help to quantify the value of high-resolution WRF ensemble forecasts *versus* global forecasts in the region.

ACKNOWLEDGEMENTS

Mailiu Diáz is grateful for the support of the National Commission for Scientific and Technological Research (CONICYT) of Chile under Grant No. 21150227. Sándor Baran acknowledges the support of the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and the National Research, Development and Innovation Office under Grant No. NN125679. Orietta Nicolis and Julio César Marín are partially supported by the Interdisciplinary Center of Atmospheric and Astro-Statistical Studies. Powered@NLHPC: this research was partially supported by the supercomputing infrastructure of the National Laboratory for High Performing Computer (NLHPC) (ECM-02). Last but not least the authors are very grateful to the reviewers for their valuable comments.

ORCID

Sándor Baran https://orcid.org/0000-0003-1035-004X

REFERENCES

Baran, S. (2014) Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics and Data Analysis*, 75, 227–238. https://doi.org/10.1016/j.csda.2014.02.013.

Baran, S., Horányi, A. and Nemoda, D. (2014) Probabilistic temperature forecasting with statistical calibration in Hungary. *Meteorology and Atmospheric Physics*, 124, 129–142. https://doi.org/10.1007/s00703-014-0314-8.

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. https://doi.org/10.1002/qj.2521.

Baran, S. and Lerch, S. (2018) Combining predictive distributions for statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34, 477–496. https://doi.org/10.1016/j.ijforecast.2018.01.005.

Baran, S. and Nemoda, D. (2016) Censored and shifted gamma distribution based EMOS model for probabilistic quantitative

precipitation forecasting. *Environmetrics*, 27, 280–292. https://doi.org/10.1002/env.2391.

Bauer, P., Thorpe, A. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55. https://doi.org/10.1038/nature14956.

Buizza, R., Houtekamer, P.L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133, 1076–1097. https://doi.org/10.1175/MWR2905.1.

Burger, F., Brock, B. and Montecinos, A. (2018) Seasonal and elevational contrasts in temperature trends in Central Chile between 1979 and 2015. *Global and Planetary Change*, 162, 136–147. https://doi.org/10.1016/j.gloplacha.2018.01.005.

Chen, F. and Dudhia, J. (2001) Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Monthly Weather Review*, 129, 569–585. https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Cortés, L. and Curé, M. (2011) Validation of the vertical profiles of three meteorological models using radiosondes from Antofagasta, Paranal and Llano de Chajnantor. In: Curé, M. , Otárola, A., Marín, J. and Sarazin, M. (Eds.) *Astronomical Site Testing Data in Chile. Revista Mexicana de Astronomía y Astrofísica (Serie de Conferencias)*, Vol. 41, pp. 64–67. Mexico: Instituto de Astronomía.

Diebold, F.X. and Mariano, R.S. (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–263. https://doi.org/10.1080/07350015.1995.10524599.

Dudhia J. (1996). A multi-layer soil temperature model for MM5. In: *Sixth PSU/NCAR Mesoscale Model Users' Workshop, Boulder, CO, 22–24 July 1996*, pp. 49–50.

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

Fraley, C., Raftery, A.E., Gneiting, T., Sloughter, J.M. and Berrocal, V.J. (2011) Probabilistic weather forecasting in R. *R Journal*, 3, 55–63.

Gebhardt, C., Theis, S.E., Paulat, M. and Ben Bouallègue, Z. (2011) Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research*, 100, 168–177. https://doi.org/10.1016/j.atmosres.2010.12.008.

Gneiting, T. (2011) Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762. https://doi.org/10.1198/jasa.2011.r10138.

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

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

Gneiting, T., Raftery, A.E., Westveld, A.H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118. https://doi.org/10.1175/MWR2904.1.

González, S. and Garreaud, R. (2019) Spatial variability of near-surface temperature over the coastal mountains in southern Chile (38 °S). *Meteorology and Atmospheric Physics*, 131, 89–104. https://doi.org/10.1007/s00703-017-0555-4.

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.

Hong, S.-Y., Dudhia, J. and Chen, S.-H. (2004) A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Monthly Weather Review*, 132, 103–120. https://doi.org/10.1175/1520-0493(2004)132<0103:ARATIM>2.0.CO;2.

Hong, S.-Y., Noh, Y. and Dudhia, J. (2006) A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, 134, 2318–2341. https://doi.org/10.1175/MWR3199.1.

Iacono, M.J., Delamere, J.S., Mlawer, E.J., Shephard, M.W., Clough, S.A. and Collins, W.D. (2005) Radiative forcing by long-lived greenhouse gases: calculations with the AER radiative transfer models. *Journal of Geophysical Research*, 113, D13103.

Janjić, Z.I. (1994) The Step-Mountain eta coordinate model: further developments of the convection, viscous sublayer, and turbulence closure schemes. *Monthly Weather Review*, 122, 927–945. https://doi.org/10.1175/1520-0493(1994)122LTHEXA0927:TSMECM>2.0.CO;2.

Kain, J.S. (2004) The Kain–Fritsch convective parameterization: an update. *Journal of Applied Meteorology*, 43, 170–181. https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2.

Lerch, S. and Baran, S. (2017) Similarity-based semi-local estimation of EMOS models. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 66, 29–51. https://doi.org/10.1111/rssc.12153.

Leutbecher, M. and Palmer, T.N. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014.

Marín, J.C., Pozo, D., Mlawer, E., Turner, D. and Curé, M. (2013) Dynamics of local circulations in mountainous terrain during the RHUBC-II project. *Monthly Weather Review*, 141, 3641–3656. https://doi.org/10.1175/MWR-D-12-00245.1.

Massey, J.D., Steenburgh, W.J., Knievel, J.C. and Cheng, W.Y.Y. (2016) Regional soil moisture biases and their influence on WRF model temperature forecasts over the Intermountain West. *Weather and Forecasting*, 31, 197–216. https://doi.org/10.1175/WAF-D-15-0073.1.

McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York, NY: Wiley.

Möller A, Thorarinsdottir TL, Lenkoski A, Gneiting T. (2015) Spatially adaptive, Bayesian estimation for probabilistic temperature forecasts. Available at: http://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. https://doi.org/10.1002/qj.49712252905.

Nakanishi, M. and Niino, H. (2006) An improved Mellor–Yamada level-3 model: its numerical stability and application to a regional prediction of advection fog. *Boundary-Layer Meteorology*, 119, 397–407. https://doi.org/10.1007/s10546-005-9030-8.

National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce. (2015) *NCEP GDAS/FNL 0.25 Degree Global Tropospheric Analyses and Forecast Grids*. Research Data Archive at the National Center for

Atmospheric Research, Computational and Information Systems Laboratory (Updated daily). Available at: https://doi.org/10.5065/D65Q4T4Z [Accessed 7th May 2019].

Pinson, P. and Hagedorn, R. (2012) Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorological Applications*, 19, 484–500. https://doi.org/10.1002/met.283.

Pleim, J.E. and Xiu, A. (2003) Development of a land surface model. Part II: Data assimilation. *Journal of Applied Meteorology*, 42, 1811–1822. https://doi.org/10.1175/1520-0450(2003)042<1811:DOALSM>2.0.CO;2.

Powers, J.G., Klemp, J.G., Skamarock, W.S., Davis, C.A., Dudhia, J., Gill, D.O., Coen, J.L., Gochis, D.J., Ahmadov, R., Peckham, S.E., Grell, G.A., Michalakes, J., Trahan, S., Benjamin, S.G., Alexander, C.R., Dimego, G.J., Wang, W., Schwartz, C.S., Romine, G.S., Liu, Z., Snyder, C., Chen, F., Barlage, M.J., Yu, W. and Duda, M.G. (2017) The weather research and forecasting model: overview, system efforts, and future directions. *Bulletin of the American Meteorological Society*, 98, 1717–1737. https://doi.org/10.1175/BAMS-D-15-00308.1.

Pozo, D., Marín, J.C., Illanes, L., Curé, M. and Rabanus, D. (2016) Validation of WRF forecasts for the Chajnantor region. *Monthly Notices of the Royal Astronomical Society*, 459, 419–426. https://doi.org/10.1093/mnras/stw600.

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

Ruiz, J.J., Saulo, C. and Nogues-Paegle, J. (2010) WRF model sensitivity to choice of parameterization over South America: validation against surface variables. *Monthly Weather Review*, 138, 3342–3355. https://doi.org/10.1175/2010MWR3358.1.

Saide, P.E., Carmichael, G.R., Spak, S.N., Gallardo, L., Osses, A.E., Mena-Carrasco, M.A. and Pagowski, M. (2011) Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRFChem CO tracer model. *Atmospheric Environment*, 45, 2769–2780. https://doi.org/10.1016/j.atmosenv.2011.02.001.

Schefzik, R. (2016a) A similarity-based implementation of the Schaake shuffle. *Monthly Weather Review*, 144, 1909–1921. https://doi.org/10.1175/MWR-D-15-0227.1.

Schefzik, R. (2016b) Combining parametric low-dimensional ensemble postprocessing with reordering methods. *Quarterly Journal of the Royal Meteorological Society*, 142, 2463–2477. https://doi.org/10.1002/qj.2839.

Scheuerer, M. (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086–1096. https://doi.org/10.1002/qj.2183.

Scheuerer, M. and Hamill, T.M. (2015) Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143, 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1.

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda M, Huang X., Wang, W., Powers, J.G. (2008) *A description of the Advanced Research WRF Version 3*. NCAR Technical Note, NCAR/TN475+ST. Boulder: National Center for Atmospheric Research.

Sloughter, J.M., Gneiting, T. and Raftery, A.E. (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, 105, 25–35. https://doi.org/10.1198/jasa.2009.ap08615.

Sloughter, J.M., Raftery, A.E., Gneiting, T. and Fraley, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135, 3209–3220. https://doi.org/10.1175/MWR3441.1.

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*, 173, 371–388. https://doi.org/10.1111/j.1467-985X.2009.00616.x.

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

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*, 3rd edition. Amsterdam: Elsevier.

Williams, R.M., Ferro, C.A.T. and Kwasniok, F. (2014) A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140, 1112–1120. https://doi.org/10.1002/qj.2198.

Zhang, Y.L., Li, X., Cheng, G.D., Jin, H.J., Yang, D.W., Flerchinger, G.N., Chang, X.L., Wang, X. and Liang, J. (2018) Influences of topographic shadows on the thermal and hydrological processes in a cold region mountainous watershed in northwest China. *Journal of Advances in Modeling Earth Systems*, 10, 1439–1457. https://doi.org/10.1029/2017MS001264.

How to cite this article: Díaz M, Nicolis O, Marín JC, Baran S. Statistical post-processing of ensemble forecasts of temperature in Santiago de Chile. *Meteorol Appl*. 2020;27:e1818. https://doi.org/10.1002/met.1818