Adv. Sci. Res., 14, 123–129, 2017
www.adv-sci-res.net/14/123/2017/
doi:10.5194/asr-14-123-2017

Advances in Science & Research
Open Access Proceedings

Multimodel probabilistic prediction of 2 m-temperature anomalies on the monthly timescale

Alfonso Ferrone, Daniele Mastrangelo, and Piero Malguzzi

Institute of Atmospheric Sciences and Climate (CNR-ISAC), 40129 Bologna, Italy

Correspondence to: Daniele Mastrangelo (d.mastrangelo@isac.cnr.it)

Received: 14 January 2017 – Revised: 12 April 2017 – Accepted: 13 April 2017 – Published: 8 May 2017

Abstract. The 2 m-temperature anomalies from the reforecasts of the CNR-ISAC and ECMWF monthly prediction systems have been combined in a multimodel super-ensemble. Tercile probability predictions obtained from the multimodel have been constructed using direct model outputs (DMO) and model output statistics (MOS), like logistic and nonhomogeneous Gaussian regression, for the 1990–2010 winter seasons. Verification with ERA-Interim reanalyses indicates that logistic regression gives the best results in terms of ranked probability skill scores (RPSS) and reliability diagrams for low–medium forecast probabilities. Also, it is argued that the logistic regression would not yield further improvements if a larger dataset was used.

1 Introduction

Multimodel ensemble forecasting has been proven to be successful on different space-time scales (e.g. Casanova and Ahrens, 2009). Using both a synthetic forecast generator and a seasonal forecast dataset, Weigel et al. (2008) showed that multimodel combination reduces overconfidence, i.e. ensemble spread is widened while average ensemble-mean error is reduced, implying a net gain in prediction skill. One of the issues of the sub-seasonal to seasonal prediction (S2S) research project, whose goal is to improve forecast skill and understanding on this timescale (Vitart et al., 2017), is the assessment of the benefits of a multimodel forecast, and how it can be constructed and implemented. The S2S database provides a set of reforecasts over a past period for model calibration purposes. Each model participating to the S2S database is furnished by a large number of reforecasts collected on many initialization dates with a typically small number of ensemble members. Then, it arises the question of how to exploit at best these reforecast sets in order to improve the skill of multimodel combinations, in particular for what concerns probability predictions.

In this work, we address this question by constructing a multimodel combination between two of the S2S models, namely the ECMWF-IFS and the CNR-ISAC monthly forecasting systems. When dealing with a small number of ensemble members as in the present case, it must be taken into account that model output statistics (MOS) based on ensemble mean quantities, and calibrated over a long retrospective dataset of reforecasts, can provide skilful probabilistic predictions that may be competitive with those obtained by direct model output (DMO) counting algorithms. Whitaker et al. (2006) give a clear example of MOS techniques applied to multimodel reforecast datasets obtained from ECMWF-IFS and NCEP-GFS global models. The main purpose of this work is then to assess which techniques work best in this particular context.

The paper is organized as follows. In Sect. 2, we describe the datasets. In this study, we limit our attention to the 2 m-temperature prediction for the winter season. In Sect. 3, we illustrate how the two-model ensemble is devised. Rather than constructing a poor man or simple multimodel, as customary in similar studies, we optimally combine the two models in a super-ensemble by computing the (unconstrained) weighting coefficients via a linear regression. In Sects. 4 and 5 we compare the skill of different techniques used to predict the probability that the 2 m temperature falls in a given tercile. In Sect. 6, by means of learning curves, we determine the convergence of the results. Conclusions are presented in Sect. 7.

2 Datasets

The reforecast simulations of the CNR-ISAC monthly prediction system, as originally created for calibration purposes in the S2S database (Vitart et al., 2017), constitute a "fixed" dataset. It covers the 30-year period ranging from 1981 to 2010, with initialization dates every 5 days from 1 January to 27 December of each year (in leap years, a 6-day gap between 25 February and 2 March is used). One simulation is obtained initializing the GLOBO model (Malguzzi et al., 2011) with ERA-Interim (Dee et al., 2011) reanalysis data at 00:00 UTC of each date. For this study, the reforecast dataset has been enlarged to a 5-member lagged ensemble in which the 4 new initial conditions are taken, from ERA-Interim, at the initialization date plus or minus 6 and 12 h. We restrict the analysis to the Northern Hemisphere winter season by choosing initialization dates in the months of December, January, and February.

The ECMWF-IFS monthly forecasting system currently produces reforecasts "on the fly" covering the past twenty years twice a week, every Monday and Thursday. For this study, we consider the 5-member ensemble generated with the ECMWF-IFS up to version CY40R1 (e.g. Vitart, 2014). Reforecasts of 2 m temperatures are downloaded from the MARS archive at the resolution of $1.5° \times 1.5°$.

By focussing on the winter season only, we can find 268 common initialization dates between the GLOBO and IFS reforecasts for the years between 1990 and 2011. However, the 1990–1991, 1991–1992, and 2010–2011 winter seasons contain few common dates, so they are excluded from the subsequent analysis, leaving a total of 258 cases distributed in 18 winters (from a minimum of 10 to a maximum of 15 cases per winter).

In both datasets, the 2 m temperature is taken at 00:00 and 12:00 UT for a forecast period of 28 days, and then averaged over the leading weeks from 1 to 4. The same average is computed on ERA-Interim reanalysis of 2 m temperature on the $1.5° \times 1.5°$ grid, which is used for regression and verification purposes.

3 Multimodel super-ensemble

In order to compute the multimodel anomaly fields, some preliminary operations have to be performed. We start by taking the ensemble mean of both ISAC-CNR and ECMWF-IFS models and the anomalies with respect to the corresponding model climate, which is function of validity time. Thus, let X_1 and X_2 represent the calibrated anomalies of the ECMWF-IFS and CNR-ISAC, X_{MM} the multimodel anomalies, C_1 and C_2 the linear regression coefficients. The anomaly fields are functions of the leading week w, latitude and longitude (i, j), and initial date d, while the coefficients C_1 and C_2 are obviously independent from d. X_{MM} is therefore computed as:

$$X_{\mathrm{MM}}(w,i,j,d) = C_1(w,i,j)X_1(w,i,j,d) + C_2(w,i,j)X_2(w,i,j,d),$$

where C_1 and C_2 minimize the mean square difference between multimodel and observed anomalies, the latter estimated from ERA-Interim reanalyses (Krishnamurti et al., 2000). A similar approach is used in Whitaker et al. (2006), where linear regression is used to combine ECMWF and NCEP reforecasts.

The sum of the regression coefficients, displayed in Fig. 1, deserves some considerations. This quantity does not have a particular physical meaning, but a rule of thumb can be derived: when the sum is close to zero, the climatology is the best possible forecast given our data. Some interesting features can be seen in the maps: there are two maxima over part of Asia and the Antarctica, where the sum reach values greater than 1.5 and increases with the week number. Some other maxima, with smaller values, can be seen over the oceans near the equator, where $C_1 + C_2$ remains always around 1. In the fourth week, some areas show sum of the coefficients almost null or even negative: over the sea on the northern part of Europe (also visible in the third week), in the northern part of Alaska and over the ocean southwest of Australia. These patterns hint to a lack of predictability over these regions.

The multimodel performance can be evaluated by adopting a cross-validation approach (Wilks, 2011), which divides the whole dataset into a training and a validation portion. The dataset is split into single boreal winter seasons, each of them including December from one year and January and February from the subsequent one. All winters are used for the k-fold cross-validation with $k = 18$ (see Sect. 2): one of them is chosen as validation set while the remaining $k - 1$ constitute the training set ("leave-one-out"), on which the regression coefficients are computed. This procedure is repeated k times, choosing a different validation winter each time (Wilks, 2011).

The multimodel skill in terms of anomaly correlation (AC), computed over typical spatial regions, is shown in Table 1. The chosen regions are the extra-tropical Northern Hemisphere (NH, 20 to 90° N), the extra-tropical Southern Hemisphere (SH, 20 to 90° S), the Equatorial Belt (EB, 20° S to 20° N), and the Euro-Atlantic region (EU, 30–80° N, 20° W–60° E). The multi-model outperforms the best model in all cases. This is of course expected for a deterministic score like root mean square, which is minimized in the super-ensemble construction, but it is less obvious for AC.

4 Probabilistic forecasts

A comprehensive forecasting system for a long scale such as the sub-seasonal one should also convey probabilistic information. In this section, we introduce the techniques used to

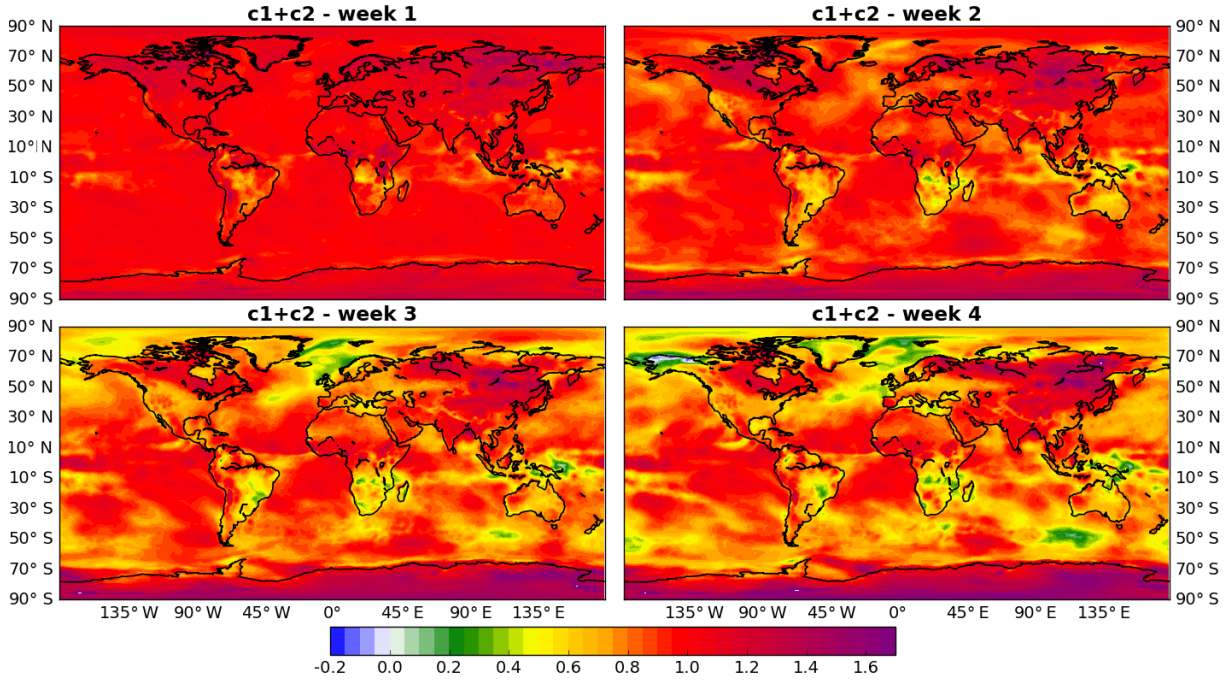

Figure 1. Sum of the regression coefficients.

Table 1. Anomaly correlation coefficient averaged over the 18 validation winters. The four pair of columns report the spatial average over the different regions described in the text. Each pair contains the results for the multimodel (MM) and the best result between the two single models (BM). The first column reports the lead week for the entire row. The value corresponding to the best performance is highlighted bold.

	NH		SH		EB		EU	
w	MM	BM	MM	BM	MM	BM	MM	BM
1	**0.93**	0.92	**0.93**	0.92	**0.90**	0.88	**0.93**	**0.93**
2	**0.68**	0.65	**0.79**	0.77	**0.77**	0.73	**0.62**	0.60
3	**0.48**	0.44	**0.75**	0.70	**0.69**	0.64	**0.38**	0.34
4	**0.44**	0.40	**0.75**	0.70	**0.67**	0.62	**0.37**	0.33

compute the predicted tercile probabilities, namely the probability that the 2 m temperature anomaly falls in each of the three regions corresponding to the first, second, and third tercile of the super-ensemble climatological distribution. We used two different groups of techniques, described in the next subsections. Example of the usage of the same techniques can be found in Hamill et al. (2004), Whitaker et al. (2006), Wilks (2006) and Wilks and Hamill (2007).

4.1 Direct Model Output (DMO)

In the direct model output (DMO) techniques, the probability estimation is based on the direct count of ensemble members that are below a given threshold. There are different ways to define the super-ensemble members (Stefanova and Krishnamurti, 2002); here, an alternative approach is proposed, in which all possible pairs of members of the two model ensembles are linearly combined with weights C_1 and C_2, obtaining a total of 25 equally reliable members.

The components of the DMO group are the Democratic Voting (DV) method and the Tukey Plotting Position (TPP). In both cases, the probability P is estimated starting from the position of the given tercile inside the ensemble. The formulas used by these two methods are $P_{\mathrm{DV1}}(T_2 \leq q) = \frac{\mathrm{rank}(q)-1}{N_{12}}$ and $P_{\mathrm{TPP1}}(T_2 \leq q) = \frac{\mathrm{rank}(q)-\frac{1}{3}}{N_{12}-1-\frac{1}{3}}$, respectively, where q denotes the tercile, T_2 the 2 m temperature anomaly, and N_{12} =25 the total number of ensemble members.

Another possibility is to define the super-ensemble members as the union of the members of each model, weighted by the corresponding regression coefficient. In this case, the DV probability estimate becomes: $P_{\mathrm{DV2}}(T_2 \leq q) = C_1 \frac{\mathrm{rank}(q)-1}{N_1} + C_2 \frac{\mathrm{rank}(q)-1}{N_2}$, where $N_1 = N_2 = 5$ and where the rank must be computed in the subset of each single model. A similar expression holds for TPP. The super-ensemble variance in this case is larger that the one computed with the previous method.

4.2 Model Output Statistics (MOS)

Logistic regression (LR) is the first model output statistic (MOS) that we consider. We test three "hypothesis functions", in which the probability is computed starting from dif-

Table 2. Ranked probability skill score computed with the Tukey plotting position ($RPSS^{(TPP)}$), logistic regression ($RPSS^{(LR)}$), and non-homogeneous Gaussian regression ($RPSS^{(NGR)}$).

w	$RPSS^{(TPP)}$	$RPSS^{(LR)}$	$RPSS^{(NGR)}$	w	$RPSS^{(TPP)}$	$RPSS^{(LR)}$	$RPSS^{(NGR)}$
	NH				SH		
1	0.65	0.68	0.62	1	0.64	0.68	0.6
2	0.29	0.34	0.31	2	0.32	0.37	0.34
3	0.11	0.20	0.16	3	0.17	0.25	0.22
4	0.07	0.17	0.14	4	0.11	0.20	0.18
	EB				EU		
1	0.49	0.55	0.44	1	0.67	0.69	0.63
2	0.31	0.39	0.31	2	0.21	0.27	0.23
3	0.24	0.32	0.25	3	−0.02	0.11	0.09
4	0.21	0.30	0.23	4	−0.04	0.10	0.08

ferent combination of the super-ensemble mean χ, its standard deviation σ, and their product. They are:

$$P_{\mathrm{LR1}}(T_2 \le q) = \frac{1}{1+\exp\left(-\theta_0-\theta_1\chi-\theta_q q\right)}$$

$$P_{\mathrm{LR2}}(T_2 \le q) = \frac{1}{1+\exp\left(-\theta_0-\theta_1\chi-\theta_2\sigma-\theta_q q\right)}$$

$$P_{\mathrm{LR3}}(T_2 \le q) = \frac{1}{1+\exp\left(-\theta_0-\theta_1\chi-\theta_2\chi\sigma-\theta_q q\right)}.$$

The terms θ_0, θ_1, θ_2 and θ_q are regression coefficients, function of the grid point and forecast week. Due to the presence of θ_q, all the hypothesis functions are in the form so called "unified logistic regression" (Wilks, 2009). This provides consistent probabilities for the three regions in which the distribution function is divided by the terciles.

The other MOS that we test is the nonhomogeneous Gaussian regression, which is defined as:

$$P_{\mathrm{NGR}}(T_2 \le q) = \Phi\left[\frac{q-(\theta_0+\theta_1\chi)}{(\theta_2+\theta_3\chi)^{\frac{1}{2}}}\right]$$

where Φ is the cumulative distribution function of the standard Gaussian distribution. The terms θ_0, θ_1, θ_2 and θ_3 are again regression coefficients.

5 Verification of probabilistic forecasts

The aim here is to evaluate the skill of the probability hindcast defined above. For the calibration and verification of the probability hindcast using MOS techniques we adopt the same cross-validation approach used in Sect. 3. The Ranked Probability Skill Score (RPSS) is computed as a "single-value estimate" for all terciles. Then, we examine the reliability diagrams for the lower tercile in order to analyse the full joint distribution of forecasts and verifying reanalysis. Example of the usage of these verifying tools in the can be found in Hamill et al. (2004), Wilks (2006), Wilks and Hamill (2007). A full explanation is presented in Wilks (2011).

Table 2 reports the values of RPSS obtained in three cases. For the DMO group, only TPP s reported because it generally gives slightly higher scores than DV. For the MOS group, LR1 (logistic regression with the super-ensemble mean as the only predictor) and NGR are considered. As it can be easily seen from Table 2, LR1 shows better performances in all the four chosen regions and for all forecast periods. It is interesting to notice that over the Euro-Atlantic region, while the DMO techniques give us almost no predictability after the second week, the use of LR1 improves significantly the skill that can be extracted from the multimodel.

In addition to RPSS, we show the reliability diagrams for the lower tercile in Fig. 2 where, for brevity purposes, we condensed in a single set of graphs, and for the whole globe, the diagrams obtained using DV, TPP, LR1, and NGR; very similar results are obtained for the upper tercile (not shown). Overall, the multimodel probabilistic prediction overforecasts the observed occurrences, especially in the first two weeks for medium–high probabilities. In the remaining weeks, the overforcasting tendency is reduced for medium probabilities, while underforecasting appears with DMO techniques for the lowest probability, which is also the most frequent as indicated in the bar graphs showing the refinement distributions.

The advantage of the MOS techniques is evident along the whole forecasting range for all the probabilities except the highest one (0.8–1.0). In the first week, LR1 (the red curve) is the closest to the bisector of the quadrant, improving the low reliability of the medium forecast probabilities. In the remaining weeks, the MOS techniques reduce the lowest probability frequency (see the refinement distributions), improving the resolution of the forecast distributions and their reliability for low–medium forecast probabilities. In particular, the underforecasting resulting from the DMO techniques for the lowest probability is completely removed. Between the

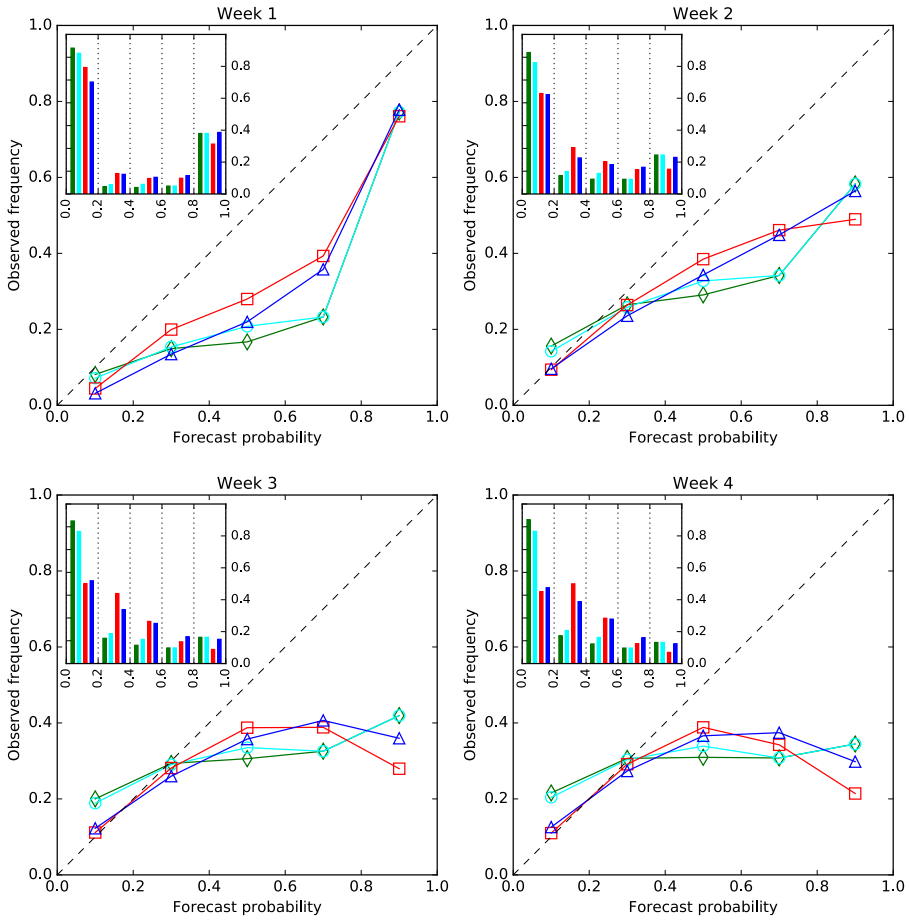

Figure 2. Reliability diagrams for the four methods (see text) applied to the lower tercile multimodel prediction. The four panels refer to the forecast weeks (average over the whole globe and over the 18 validation winters). Democratic voting (DV): green; Tukey plotting position (TPP): cyan; logistic regression (LR): red; and nonhomogeneous Gaussian regression (NGR): blue. On the x axis of each graph there are the five probability intervals, while on the y axis the conditional probabilities. In the refinement distribution (the bar graph), the frequency of each probability interval is reported. The no-resolution threshold, given by the averaged observed frequency, is 0.336.

two MOS techniques, LR1 systematically outperforms NGR for the 0.0–0.6 probability range.

For the highest probability value considered, DMO methods outperform MOS techniques. Specifically, in the extended range, LR1 seems unable to produce high-probability forecasts, as evident from the strong reduction of the associated forecast frequency.

In summary, LR1 drastically improves the low probability forecasts, slightly improves medium probability forecasts and is unable to provide skilful high probability forecasts.

6 Learning curve

After having shown that the LR technique provides the best performance, except for the high-probability forecasts, our aim in this section is to determine the possible occurrence of over-fitting or under-fitting, and therefore the potential for improving this particular technique. The method of the learning curve consists in repeating the computation of LR coefficients for different dimensions of the training set (10, 20, 30, 40, 50, 100, 150, and 200 elements). With the recomputed LR coefficients, we evaluate the RPSS averaged over the globe for each of the four forecast weeks. The results are summarized in Fig. 3. As expected, for small training sets the algorithm performs exceptionally well in the training phase, while in the validation mode it shows poor results. The two values become closer for increasing training set dimensions and, above the threshold of 100 elements, the scores remain nearly constant. This is a definite indication that the LR has extracted all the available information from the data. However, it cannot be excluded that there actually is some additional information that LR is not capable to catch or, in other words, that LR is under-fitting. In this case, the results do not improve by adding more training elements, and different solution should be adopted, like using more complex fittings.

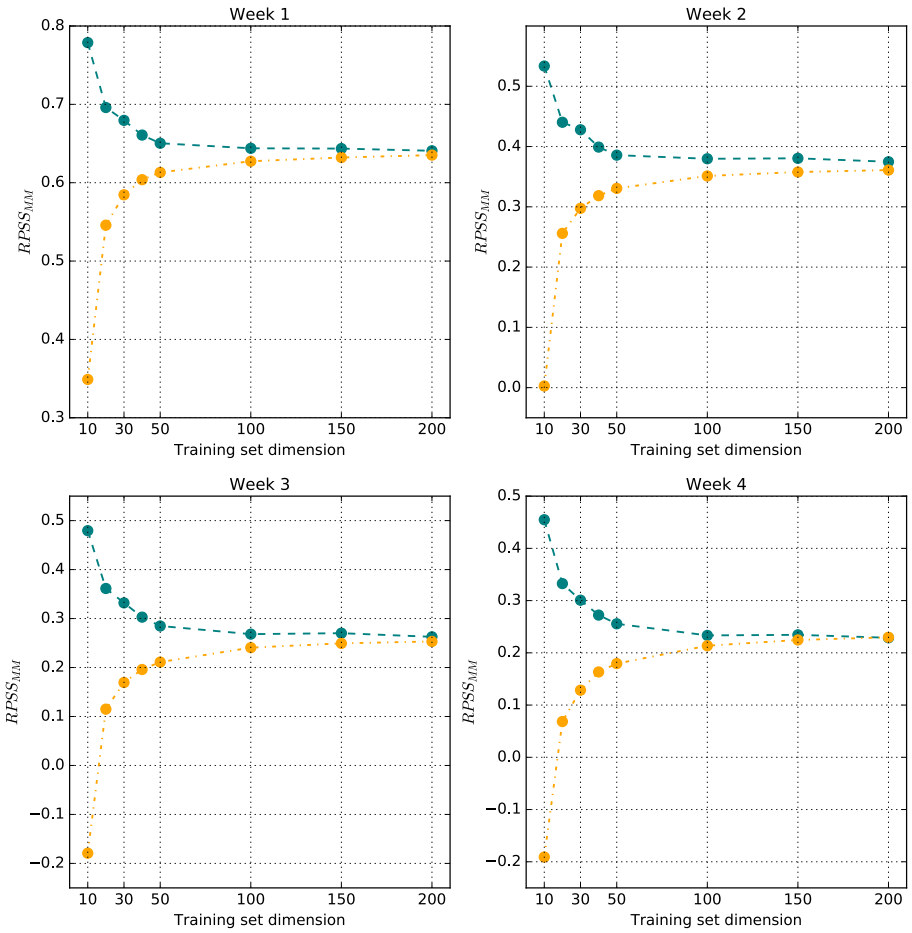

Figure 3. Learning curve for the LR (using the ensemble mean only) for weeks from 1 to 4. The y axis reports the ranked probability skill score averaged over the whole globe. On the x axis it is reported the dimension of the reduced training set. The scores obtained on the training (validation) sets are shown in teal (orange).

7 Conclusions

The aim of this study is the evaluation of the predictive skill of the super-ensemble obtained by combining through linear regression the reforecasts of the ECMWF-IFS and CNR-ISAC monthly prediction systems. The main focus is, in particular, on probabilistic wintertime 2 m temperature predictions: we compare different techniques for extracting forecast tercile probabilities from the multimodel ensemble, namely direct model output (DMO) and model output statistics (MOS). We conclude that the logistic regression based on the ensemble mean as only predictor gives the best performance in terms of RPSS and reliability diagrams for low–medium forecast probabilities.

We finally analyse the behaviour of logistic regression with different dimensions of the training dataset. We conclude that no further improvements are possible from an extension of the dataset, unless a more complex algorithm is used.

The results here presented depend on the rather limited number of ensemble members at our disposal. Conclusions may be quite different if more members are added to each single model, or if more models are considered. However, the outcomes obtained in the rather simple framework described here suggest that the multimodel ensemble could be easily implemented to improve CNR-ISAC operational monthly forecasts.

Data availability. The ECMWF reforecasts and ERA-Interim reanalyses were downloaded, after log in, through the MARS catalogue http://apps.ecmwf.int/mars-catalogue/.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors wish to thank an anonymous reviewer for the useful suggestions that helped to improve the manuscript. This work was partly supported by the Italian Civil Protection Agency.

Edited by: Á. G. Muñoz
Reviewed by: two anonymous referees

References

Casanova, S. and Ahrens, B.: On the Weighting of Multimodel Ensembles in Seasonal and Short-Range Weather Forecasting, Mon. Weather Rev., 137, 3811–3822, doi:10.1175/2009MWR2893.1, 2009.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteorol. Soc., 137, 553–597, 2011.

Hamill, T. M., Whitaker, J. S., and Wei, X.: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts, Mon. Weather Rev., 132, 1434–1447, 2004.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., and Surendran, S.: Multimodel Ensemble Forecasts for Weather and Seasonal Climate, J. Climate, 13, 4196–4216, doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2, 2000.

Malguzzi, P., Buzzi, A., and Drofa, O.: The meteorological global model GLOBO at the ISAC-CNR of Italy: Assessment of 1.5 years of experimental use for medium range weather forecast, Weather Forecast., 26, 1045–1055, 2011.

Stefanova, L. L. and Krishnamurti, T. N.: Interpretation of Seasonal Climate Forecast Using Brier Skill Score, The Florida State University Superensemble, and the AMIP-I Dataset, J. Climate, 15, 537–544, doi:10.1175/1520-0442(2002)015<0537:IOSCFU>2.0.CO;2, 2002.

Vitart, F.: Monthly forecasting at ECMWF, Mon. Weather Rev., 132, 2761–2779, 2004.

Vitart, F.: Evolution of ECMWF sub-seasonal forecast skill scores, Q. J. Roy. Meteorol. Soc., 140, 1889–1899, doi:10.1002/qj.2256, 2014.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H. S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D. J., Xiao, H., Zaripov, R., Zhang, L.: The Sub-seasonal to Seasonal Prediction (S2S) Project Database, B. Am. Meteorol. Soc., 98, 163–173, doi:10.1175/BAMS-D-16-0017.1, 2017.

Weigel, A. P., Liniger, M. A., and Appenzeller, C.: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, Q. J. Roy. Meteorol. Soc., 134, 241–260, 2008.

Whitaker, J. S., Wei, X., and Vitart, F.: Improving week-2 forecasts with multimodel reforecast ensembles, Mon. Weather Rev., 134, 2279–2284, 2006.

Wilks, D. S.: Comparison of ensemble-MOS methods in the Lorenz96 setting, Meteorol. Appl., 13, 243–256, 2006.

Wilks, D. S.: Extending logistic regression to provide full-probability-distribution MOS forecasts, Meteorol. Appl., 16, 361–368, 2009.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 3rd Edn., Academic Press, Oxford, UK, 2011.

Wilks, D. S. and Hamill, T. M.: Comparison of ensemble-MOS methods using GFS reforecasts, Mon. Weather Rev., 135, 2379–2390, 2007.