ENSEMBLES stream-1 hindcasts: from the season to the decade with four coupled models

Michel Déqué

Météo-France/CNRM, CNRS/GAME
Toulouse, France
deque@meteo.fr

ABSTRACT

Preliminary results of ENSEMBLES seasonal to decadal hindcasts are presented. They are based on stream-1 experiment which covers a shorter period than DEMETER. Three methods of generating ensembles with alarge enough spread are presented: the traditional multimodel approach, the stochastic physics and the perturbed parameters. The skill and spread of northern hemisphere height are examined. Given the sample uncertainty, it is difficult to discriminate the three methods. Decadal skill is also examined: nothing is found for northern hemisphere height, but the impact of greenhouse gas warming is clearly identified.

1 Introduction

The concept of multimodel seasonal forecast was born in Europe with the PROVOST project (e.g. Doblas-Reyes et al.,2000). In this project, 3 models (ECMWF, Met Office and Météo-France) contributed to a hindcast system based on ERA15 reanalysis. As the models did not incorporate any ocean component, the predictability was deliberately overestimated (monthly observed sea surface temperature -sst- was used). The following multimodel project, DEMETER (Palmer et al., 2004) evaluated the actual seasonal predictability by a simple assimilation scheme for the ocean and coupled ocean-atmosphere models. The project was based on ERA40 reanalysis, so the longer hindcast period (three times that of PROVOST) improved the statistical significance of the scores, in particular at mid latitudes. The atmosphere models were improved versions of those used in PROVOST, and a fourth atmosphere model, originating at Max Plank Institute, was added. By combining versions of the 4 atmosphere models with versions of 3 ocean models, 7 coupled model were used in the project. As in PROVOST, the results showed that the average of the different models was at least as good as the best model (which varies according to the criterion) as far as deterministic skill is concerned. When considering probability forecasts, the multimodel was clearly superior to any single model, because an ensemble based on the perturbation of initial conditions takes into account the unpredictability due non-linearitiy of the system (the so-called butterfly effect) but ignores the uncertainty due to model imperfection. At the end of the project, three centres (ECMWF, Met Office and Météo-France) decided to continue the experiment in real time, with regular updates of their models. This multimodel forecast is named EUROSIP.

The ENSEMBLES European project covers many aspects of ocean/atmosphere modeling, among which seasonal to decadal prediction. One of the goals of this project is to update and extend DEMETER. As this is a 5-year project, it was decided to start with a draft experiment, namely stream-1, in which modelers can test their models and methods, and impact users will train their application models. The present study is devoted to results of this stream-1, with a focus on midlatitude circulation in the northern hemisphere. In Section 2 we examine the monthly and seasonal skill of three ensembles. The spread of these ensembles is evaluated in Section 3. Three out of the four models have extended two hindcasts to the decadal range. The predictability of 10-year means is addressed in Section 4. A summary and perspectives for stream-2 are given in Section 5

2 Skill of ensemble forecasts

The ENSEMBLES stream-1 exercise has been defined as follows. Four coupled models (ECMWF, Met Office, Météo-France and Institut für Meereskunde) have been used over the 1991-2001 period. Due to the availability of better ocean observations, a more refined assimilation scheme than in DEMETER was used to initialize the ocean models. Two hindcasts per year were issued: from May to November (7 months) and from November to December next year (14 months). Each model produced nine members by perturbing slightly the initial ocean state (as in DEMETER). In addition, two models produced an ensemble of 9 members which takes into account the model imperfection. The Met Office used the so-called perturbed parameters (Murphy et al., 2007), which consists of modifying at random a few uncertain coefficients of the parameterization package. ECMWF used the so-called stochastic physics (Shutts and Palmer, 2007) which consists of adding a random term to the model equations at each time step. In both methods the choice of the size of the perturbation is empirical and must be large enough to be effective, but not too large to avoid generating an unrealistic climate. In the initial perturbation method, the perturbation size is less critical since the error growth saturates during the first month of model integration. The philosophy behind stochastic physics is to add a parameterization in the model (a kind of diffusion) whereas perturbed parameters tend to create a simple multimodel as wide as one wishes.

As mentioned in Section 1, we restrict to 500 hPa height between 20°N and 80°N. Three ensemble methods are considered: the multimodel with initial perturbation technique (referred to as MM), the unimodel with perturbed parameters technique (PP), and the unimodel with stochastic physics (SP). In order to do a fair evaluation, the MM ensemble is based on 9 members (3 from ECMWF, 2 from other models). This fairness is rewarded by the fact that many 9-member subsamples can be generated by selecting each year the individual members of each model. This random selection has been repeated 500 times, and the mean, 2.5 ang 97.5 percentiles of the scores have been calculated. The 95% interval does not measure the full uncertainty about the score (if we use a different evaluation period, we may be outside this range), but only the uncertainty due to the finite size of the ensemble. If we had an infinite ensemble, the mean would be the expectation and there would be no sampling uncertainty. If we use a longer verification period (as in DEMETER or in the forthcoming stream-2) we can reasonably expect a narower interval.

Figure 1 shows the anomaly correlation for the November seasonal hindcasts. At month 1 (not shown), all methods are equivalent (ACC=0.4). The traditional DJF hindcast has a 0.25 ACC for MM whereas it is 0.20 for both PP and SP, but this is not significantly different. We can also calculate confidence levels for a pure random forecast by scrambling the verification years. These levels (not shown) indicate that the ACCs are significant at 97For month 7 (AMJ) the PP is significantly superior.

Figure 2 shows the ACC for summer hindcasts. At month 1 (not shown), the best is SP (ACC=0.52), then MM (ACC=0.45) the PP (ACC=0.37). As far as seasonal means are concerned, MM and SP are above PP, but the differences are not significant. We can remark that PP provides the same skill as when it starts 6 months earlier.

3 Spread of ensemble forecasts

If we consider probabilistic forecast skill, ACC is not enough because the spread of each ensemble has to be considered. Ideally, the spread should be similar to the error so that the verification field is compatible with the ensemble. In other words, for a large ensemble, one member should be close to observation (obviously we ignore which one a priori). The traditional measure of spread is ensemble standard deviation. However seasonal skill has a strong interannual variability. Some years are characterized by a large signal (e.g. as a response to an ENSO forcing). But this large signal may be accompanied by a large standard deviation, or at least a standard deviation as large as in other years. So a better measure for spread should take into account, for a given year, the ratio between the standard deviation and the mean. A simple index is the ACC between one member and the average of the others. If we consider year by year series of model standard deviation, it is quasi-constant for 500 hPa height, whereas intra-ensemble ACC exhibits interannual variability which could be related to some

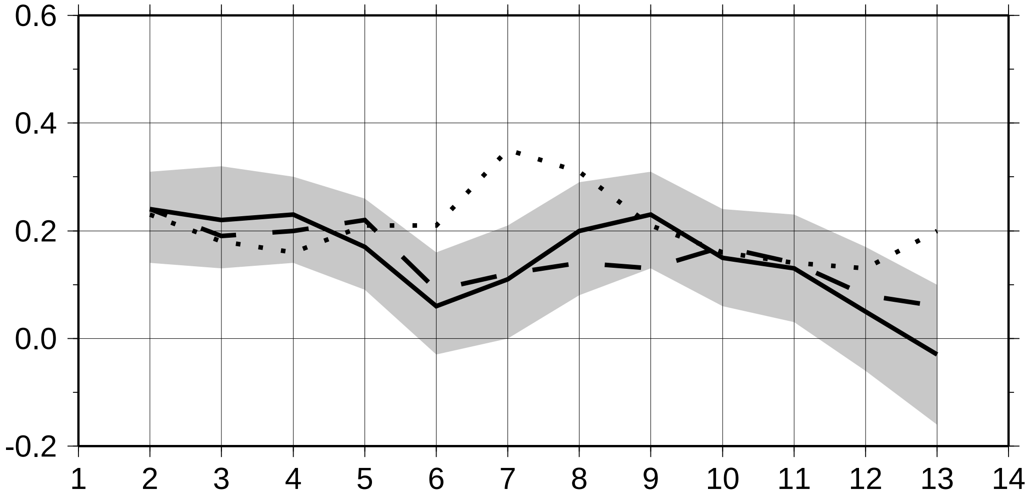

Figure 1: Anomaly correlation for the 500 hPa height hindcasts starting at 1st November for 3-month running means; multimodel (solid), perturbed parameters (dot) and stochastic physics (dash). The shaded area corresponds to 95% interval for the multimodel.

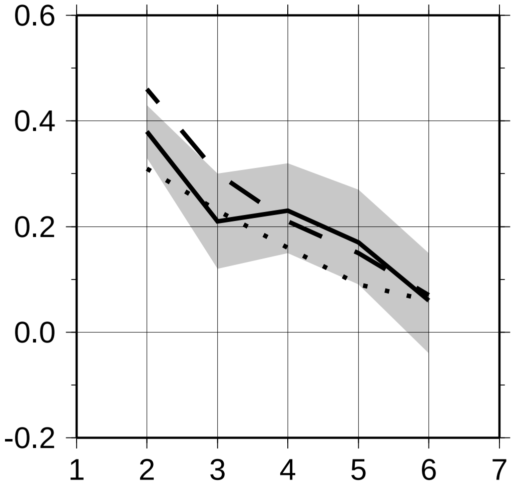

Figure 2: As Figure 1 for hincasts starting at 1st May.

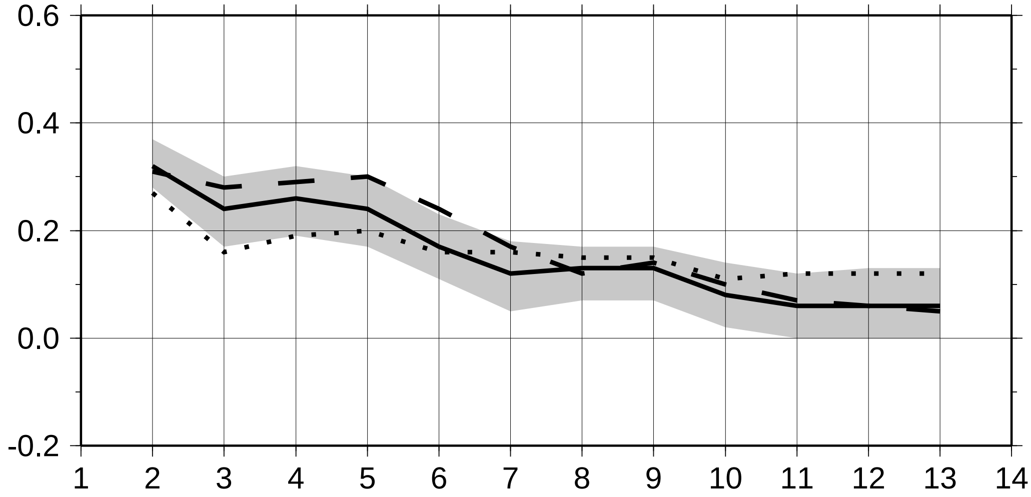

Figure 3: Ensemble spread for the 500 hPa height hindcasts starting at 1st November for 3-month running means, measured by anomaly correlation in perfect model mode; multimodel (solid), perturbed parameters (dot) and stochastic physics (dash). The shaded area corresponds to 95% interval for the multimodel.

spread-skill connection. This ACC has been stabilized by taking each member as a verification and averaging the 9 results. This approach is known as perfect model approach, although here we use several models (MM and PP). Note that each member is centered with respect to its own 11-year average. This important in the case of PP and MM, since each member has its own climate, and centering the ensemble as a whole would inflate the spread.

Figure 3 shows the spread of hindasts starting at 1st November. Here the lower the ACC, the higher the spread. It can be seen that PP provides a larger spread than the other two methods, in agreement with its lower scores till month 4. If we compare the results of Figure 3 with Figure 1 as far as DJF is concerned, the MM method provides a satisfactory spread (ACC=0.25 in both cases), SP underestimates the spread (ACC=0.30 versus 0.20) and PP overestimates the spread (ACC=0.15 versus 0.20). The overestimation of spread by PP is particularly acute in the case of MJA hindcasts (ACC=0.15 versus 0.34). If this result is confirmed by longer experiments, this means that the perturbations of the parameters are too large for seasonal forecasts. With stream-1 (11 cases) the uncertainty about a 9-member ACC is large and makes it difficult to draw definite conclusions

In summer (Figure 4) the spread is similar for the 3 methods. For JJA mean, PP is consistent, but MM and SP overestimate the spread. This result, in opposition to winter hindcasts, shows that we must be very careful in interpreting results with large sampling uncertainties.

4 Decadal predictions

Decadal predictions in Europe have started with the PREDICATE project. In perfect model approach, results have shown some potential predictability. The main problem in real decadal forecasting is a good initialization of the deep ocean which contains the only information that is kept in memory by the system beyond month 6 of a model integration. In ENSEMBLES, a particular effort has been done in this sense (Weisheimer et al., 2007). In stream-1 only two cases were considered, one for the 1965-1974 decade and one for the 1994-2003 decade. Three models have participated (ECMWF, Met Office and Météo-France), with 9 members in each case.

Figure 5 shows the differences between the two decades. In ERA40 a big negative anomaly is found in the Pacific, with positive anomalies on both sides; a stationary wave is found in the Atlantic. None of these features are found in any of the 3 models (and neither in the multimodel average, not shown). This can be explained

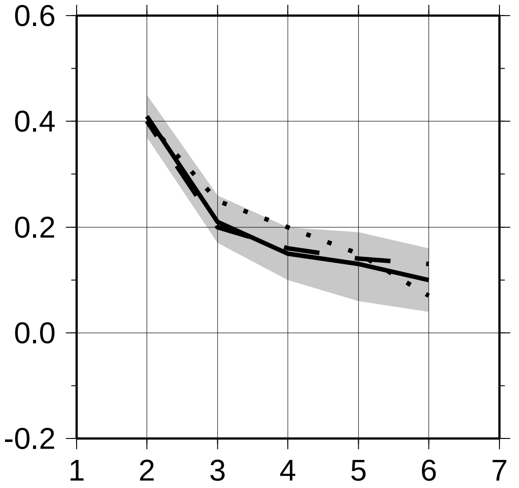

Figure 4: As Figure 3 for hincasts starting at 1st May.

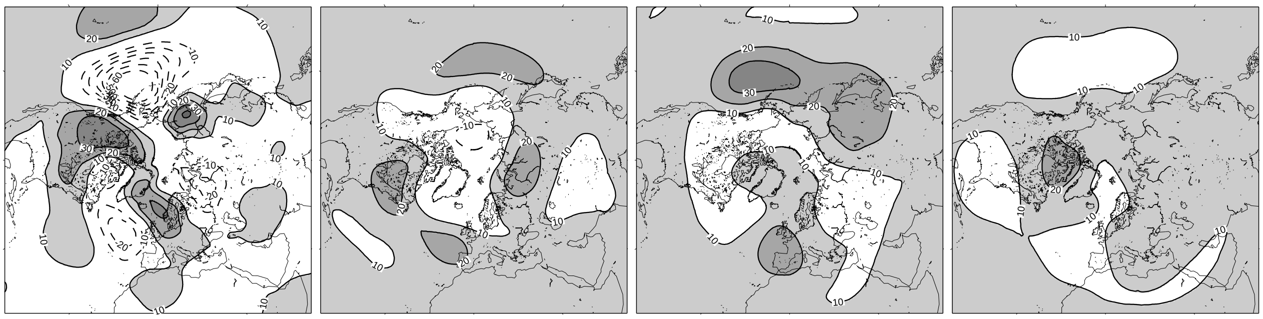

Figure 5: Difference in 500 hPa height for the mean 1994-2003 minus the mean 1965-1974; from left to right, ERA40, ECMWF, Met Office and Météo-France; contour interval 10 m, 0-contour omitted, negative contours dashed, shading above 10 m.

by the fact that in 1965 no systematic network of the ocean sub-surface, neither altimetry data was available. So the initial state of the first hindcast uses only the observed sst and possibly the tropical mixing layer depth (trough surface wind assimilation). At this stage, we have no ingredient to run something different from a climatology run.

However, another forcing is present in the simulation: the greenhouse gas concentration. Figure 6 shows the difference between the two decades for 850 hPa temperature. The spatial patterns do not coincide, but all models agree with ERA40 for a warming in response to the highly previsible (by extrapolation at 10-year range) concentration of greenhouse gases. The average for the northern hemisphere is 0.4K for ERA40, 0.5K for ECMWF and 0.6K for Met Office and Météo-France.

5 Conclusions

This study has explored two innovative aspects of stream-1. The first one is the use of different methods for generating ensembles. Due to the shortness of the sample, the different behaviors of the scores and of the scores cannot be discriminated: the MM, PP and SP methods have the same skill and spread, as far as northern midlatitudes general circulation is concerned. The second aspect concerns decadal predictability. None of the 3 models involved is able to reproduce something looking like observation. A t-test performed for each model

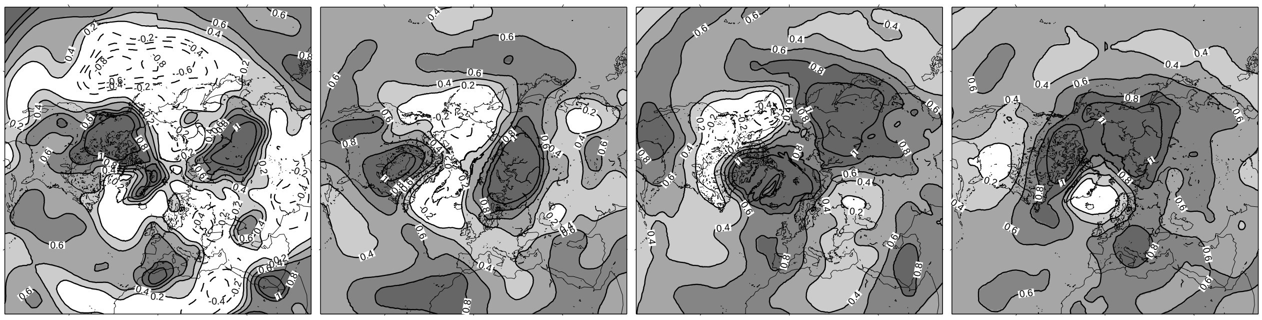

Figure 6: As Figure 5, for 850 hPa temperature; contour interval 0.2K, shading above 0.2K.

confirms that the patterns of the inter-decade differences are produced natural interannual variability North of 40°N . But there is a clear predictability of the warming between the two periods, due to increase in greenhouse gas concentration, at least at the hemispheric scale.

The perspectives of this work are to apply the same calculation to stream-2 of ENSEMBLES project. Indeed the uncertainty about the seasonal score and spread estimates is expected to be divided by about 3 (based on DEMETER estimates), so that characteristics of a particular ensemble method could be evidenced. On the other hand stream-2 will provide five independent decadal hindcasts. We cannot exclude that the failure of the pair of hindcasts we have examined is due to bad luck: if we take two seasonal hindcasts at random, we may get a similar result although we know that skill exists for the northern hemisphere at month2-month4 range (and beyond for the tropical Pacific).

Acknowledgements

The author is indebted to A. Weisheimer and F.J. Doblas Reyes for providing monthly means of the individual hindcasts. This study was partly supported by European Commission FP-6 project ENSEMBLES (GOCE-CT-2003-505539).

References

Doblas-Reyes, F.J., M. Déqué and J.Ph. Piedelievre, 2000. Model and multimodel spread in the PROVOST seasonal forecasts : application to probabilistic forecasts. Q. J. Roy. Meteor. Soc., 126, 2069-2088

Murphy, J. M., Booth, B. B. B., Collins, M., Harris, G. R., Sexton, D. and Webb, M. 2007. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Phil. Trans. R. Soc. A, 365, 1993-2028. (doi:10.1098/rsta.2007.2077)

Palmer, T.N., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Delecluse, M. Déqué, E. Diez, F. J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J.-F. Guérémy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J.-M. Terres, M. C. Thomson, 2004: Development of a European Multi-Model Ensemble System for Seasonal to Inter-Annual Prediction (DEMETER). Bull. Am. Meteorol. Soc., 85, 853–872

Shutts, G. and Palmer, T.N. 2007. Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. J. Climate, 20, 187-202.

Troccoli, A. and Palmer, T. N. 2007. Ensemble decadal prediction from analysed initial conditions. Phil. Trans. R. Soc. A 365. (doi:10.1098/rsta.2007.2079)

Weisheimer, A., F. Doblas-Reyes, P. Rogel, N. Keenlyside, M. Balmaseda, J. Murphy, D. Smith, M. Collins, B. Bhaskaran, and T. Palmer (2007). Initialisation strategies for decadal hindcasts for the 1960-2005 period within the ENSEMBLES project. ECMWF Tech. Memo., 521.