Can multi-model combination really enhance the predictive skill of probabilistic ensemble forecasts?

A.P. Weigel, M.A. Liniger and C. Appenzeller

Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland
(andreas.weigel@meteoswiss.ch)

Probabilistic forecasts with ensemble prediction systems have found a wide range of applications in weather and climate risk management, and their importance grows continuously. While "classical" ensembles account for the uncertainties in the initial conditions, the uncertainties due to model formulation can be considered in a pragmatic way by combining single model ensembles (SME) to a multi-model ensemble (MME). Indeed, it has been shown that forecasts issued on the basis of such MMEs on average outperform any single model strategy. Multi-model forecasts are even found to be superior to a "best model approach", that is a strategy which for any given forecast context (location, predictand, start time, etc.) always selects the respective best single model available. Given that a MME contains information of all participating models, including the less skillful ones, the question arises as to why, and under which conditions, a multi-model can have higher skill than if simply the best participating single model is chosen. In this contribution, we firstly seek to resolve this supposed paradox by applying a synthetic toy model, and, secondly, to quantify the gain in skill in a real multi-model seasonal forecasting system (the DEMETER data set).

The climate forecast toy model is designed such that it allows the generation of perfectly calibrated SMEs of any ensemble size and prediction skill. Additionally, the degree of "overconfidence" can be prescribed, i.e. the SMEs can be forced to have distributions which, while being sharp, are centered at the wrong value. MMEs are then constructed from weighted averages of these SMEs. As a skill measure, we apply the "debiased ranked probability skill score" (RPSSd; Müller et al., 2005; Weigel et al., 2006), which is favorable in the context of multi-model studies. The RPSSd measures the true gain in skill due to model combination while ignoring the gain in intrinsic reliability due to increased ensemble size.

Using this toy model, systematic model-combination experiments are carried out. We evaluate how multi-model performance depends on ensemble size, prediction skill and overconfidence of the participating SMEs. The central conclusion drawn from this study is that MMEs can indeed outperform a "best model approach", but only under certain conditions. We substantiate this conclusion by quantifying the gain in skill in a real multi-model seasonal forecasting system using near surface temperature forecasts from the DEMETER data set.