Sensitivity of ensemble forecast verification to model bias

Jingzhuo Wang\textsuperscript{1}, Jing Chen\textsuperscript{1}, and Jun Du\textsuperscript{2}
\textsuperscript{1}Numerical Weather Prediction Center, China Meteorological Administration, Beijing, China
\textsuperscript{2}Environmental Modeling Center, NOAA/NWS/NCEP, College Park, Maryland

This study demonstrates how model bias can adversely affect the quality assessment of an ensemble prediction system (EPS) by verification metrics. A regional EPS [Global and Regional Assimilation and Prediction Enhanced System-Regional Ensemble Prediction System (GRAPES-REPS)] was verified over a period of one month over China. Three variables (500-hPa and 2-m temperatures, and 250-hPa wind) are selected to represent "strong" and "weak" bias situations. Ensemble spread and probabilistic forecasts are compared before and after a bias correction. The results show that the conclusions drawn from ensemble verification about the EPS are dramatically different with or without model bias. This is true for both ensemble spread and probabilistic forecasts. The GRAPES-REPS is severely underdispersive before the bias correction but becomes calibrated afterward, although the improvement in the spread' spatial structure is much less; the spread-skill relation is also improved. The probabilities become much sharper and almost perfectly reliable after the bias is removed. Therefore, it is necessary to remove forecast biases before an EPS can be accurately evaluated since an EPS deals only with random error but not systematic error. Only when an EPS has no or little forecast bias, can ensemble verification metrics reliably reveal the true quality of an EPS without removing forecast bias first. An implication is that EPS developers should not be expected to introduce methods to dramatically increase ensemble spread (either by perturbation method or statistical calibration) to achieve reliability. Instead, the preferred solution is to reduce model bias through prediction system developments and to focus on the quality of spread (not the quantity of spread). Forecast products should also be produced from the debiased but not the raw ensemble.