Impact of bias on traditional precipitation verification measures and on bias adjusted scores

Keith Brill (1) and Fedor Mesinger (2)

(1) NCEP/Hydrometeorological Prediction Center, Camp Springs, MD, USA (keith.brill@noaa.gov), (2) NCEP/EMC Camp Springs, MD, USA, and SASA, Belgrade, Serbia (fedor.mesinger@noaa.gov)

The steady increase in resolution of operational NWP models has generated considerably invigorated interest in precipitation verification methods for quite some years now. This has been largely prompted by the understandable difficulty of high resolution precipitation forecasts to conform to high resolution analyses when verified on the forecast or on the analysis grid even in cases when the forecast is obviously quite successful. Remapping to a lower resolution verification grid as done for many years at NCEP reduces but does not eliminate the problems. Therefore, numerous new methods have been proposed that aim to define and focus on precipitation events and/or features of the two fields. At the same time, understanding of the behavior of conventional scores has been advanced. Thus, a critical performance ratio (CPR) has been defined which quantifies the requirement for the probability of detection (POD) to change as bias changes so that a performance measure improves at a given value of bias. This can and has been used to assess the sensitivity of various measures to "hedging", increasing or decreasing bias with the aim of improving the measure at hand. In addition, modifications of the scores have been proposed to arrive at values adjusted or corrected to unit bias. In our previous analyses, of the verification measures inspected the Clayton skill score (CSS) and the equitable threat score bias adjusted following an assumed relationship between false alarms and misses (BCETS in the notation of a recent review by Gilleland et al.) have demonstrated the greatest resistance to hedging. In this contribution we apply these two analysis tools to additional verification measures. We also include an example by evaluating the bias adjusted scores for a parallel in which the impact would seem to be of a particular interest in view of a distinct difference in the biases of the two systems compared.