



Issues of spatial forecasts, hypothesis tests, multiple comparison and field significance.

M. Pocerlich, E. Gilleland

National Center for Atmospheric Research

Boulder, CO USA

Since a skill score is the result of a function of random data, it is reasonable to ask the question, does this forecast have a positive skill (assuming positive scores mean skill) or with another sample of data could one reasonably expect a negative result? This question can be structured as a hypothesis test. Many weather forecasts are created and conveyed in a spatial format. Spatial information on the performance and limits in skill of a model is desired. In this context one is often examining many locations within the same proximity or a gridded forecast verified with a gridded analysis. Conducting tests at each site independently creates statistical issues with multiple comparisons. If enough hypothesis tests are conducted, some will appear to be significant simply by chance. In addition, both the forecasts and observations are spatially correlated so should not be treated as independent single values.

In recent years, two significant papers have been published that relate to these issues. They address the topics of field significance and the use of the false discovery rates to identify regions of statistical significance (Wilks (2006) and Ventura et al (2004)). In this talk, we show results which apply these methods to spatial forecasts and identify regions of skill as a function of forecast lead times. Temperature forecasts are evaluated using a mean squared error skill score. Confidence intervals are developed both parametrically and using re-sampling techniques.

(Reference: Ventura, V, C. Paciorek, and J. Risbey, 2004: Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data. *J. of Climate*, **17**, 4343-4356. Wilks, D.S, 2006: On “Field Significance” and the False Discovery Rate. *J. Appl. Meteor. and Climatology*, **45**, 1181-1189.)