Another look at Proper Scoring Rules

Laurence J. Wilson
Environment Canada, Science and Technology Branch
Dorval, Quebec
Tilmann Gneiting
University of Washington
Seattle, Washington

Outline

- Some history of proper scores
- Theory and experiment – localized ensemble scores
- Is properness important?

Definition of "proper"

- A proper scoring rule is a score for which a forecaster obtains a best score value by forecasting according to his/her true beliefs.
- A strictly proper scoring rule results in best score value only if the forecaster forecasts according to his/her true beliefs.

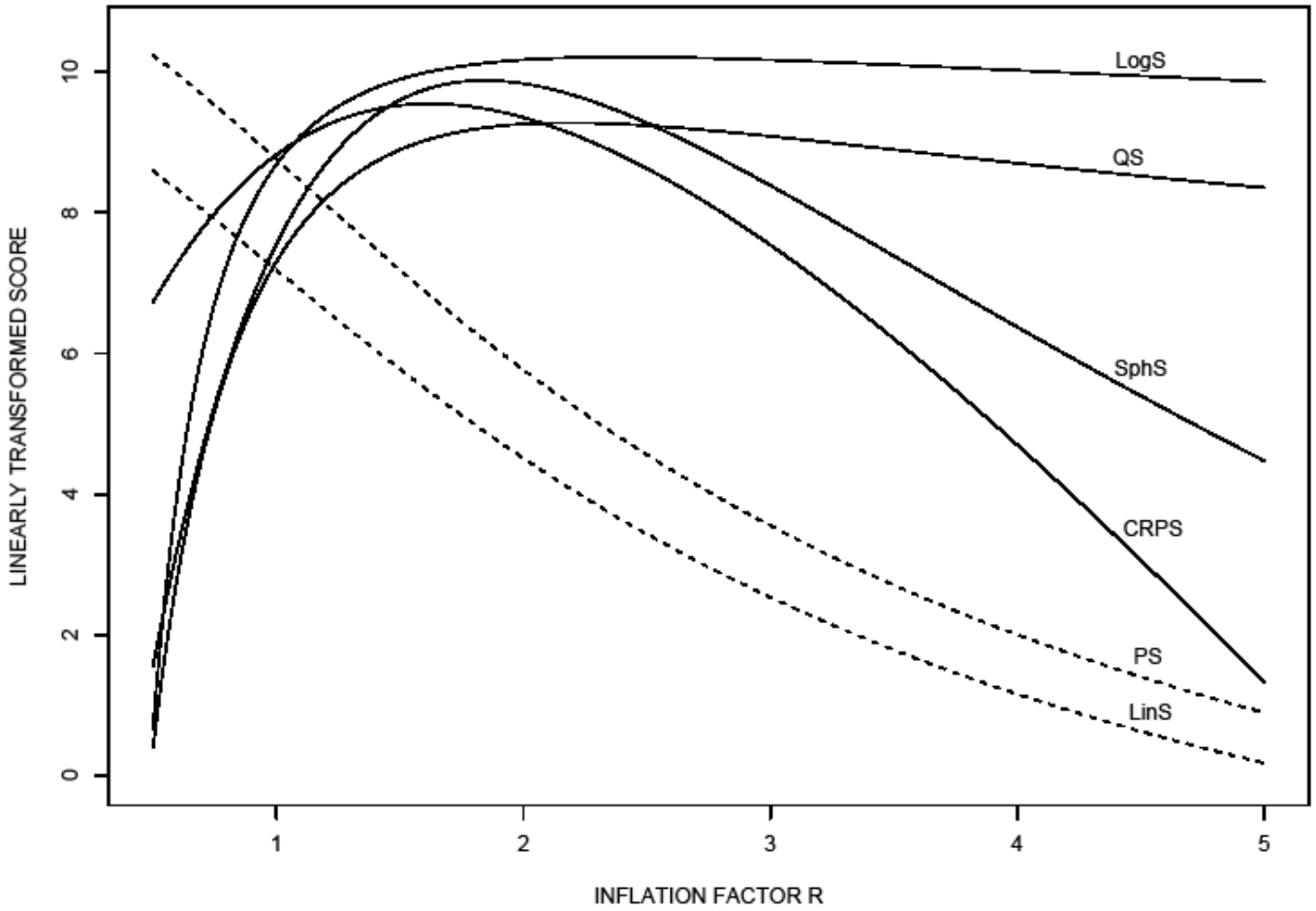
Some history

- Most scores for probability forecasts of categorical variables proved to be proper in late 1960s and early 70s by Murphy and Stael von Holstein (several papers)
  - Brier score, rank probability score
  - Ratio skill scores asymptotically proper, but often improper in the way calculated
  - Linear probability error is improper
- Recent revival of interest in proper scores for verification of ensemble distributions
  - E.g. Gneiting and Raftery, paper accepted by J. Amer Statistical Assn.
    - Comprehensive theoretical review: Scores for pdfs which are non-linear are proper, linear scores are not.

Results from Gneiting and Raftery, 07

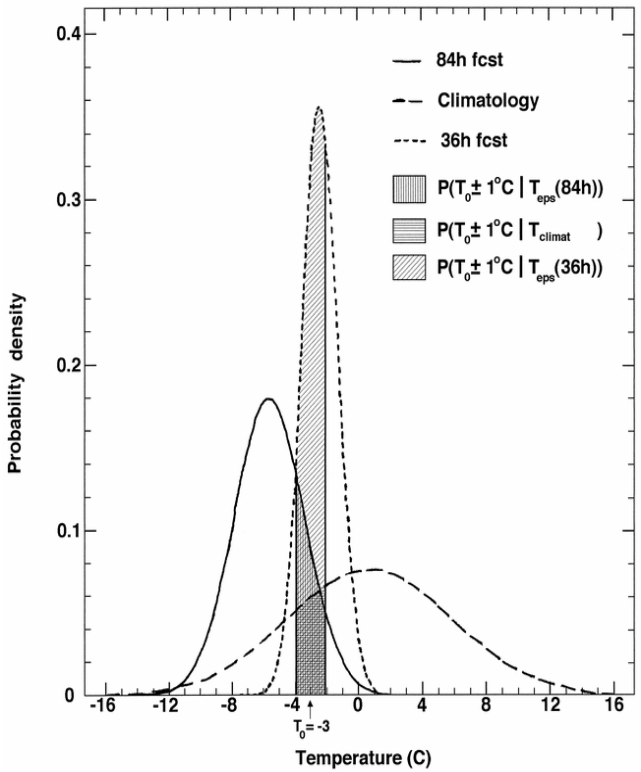
Inflation factor of standard deviation for several scores

Based on U Wash ensemble forecasts for temperature at 160 stations.



Proper Experiment

- Two scores:
  - Probability score (Wilson et al, 1999)
  - Ignorance score (Roulston and Smith, 2002)



$$P(T_{obs} \mid T_{eps}) = \int_{T-\Delta T}^{T+\Delta T} f\left(T_{eps}\right) dT$$

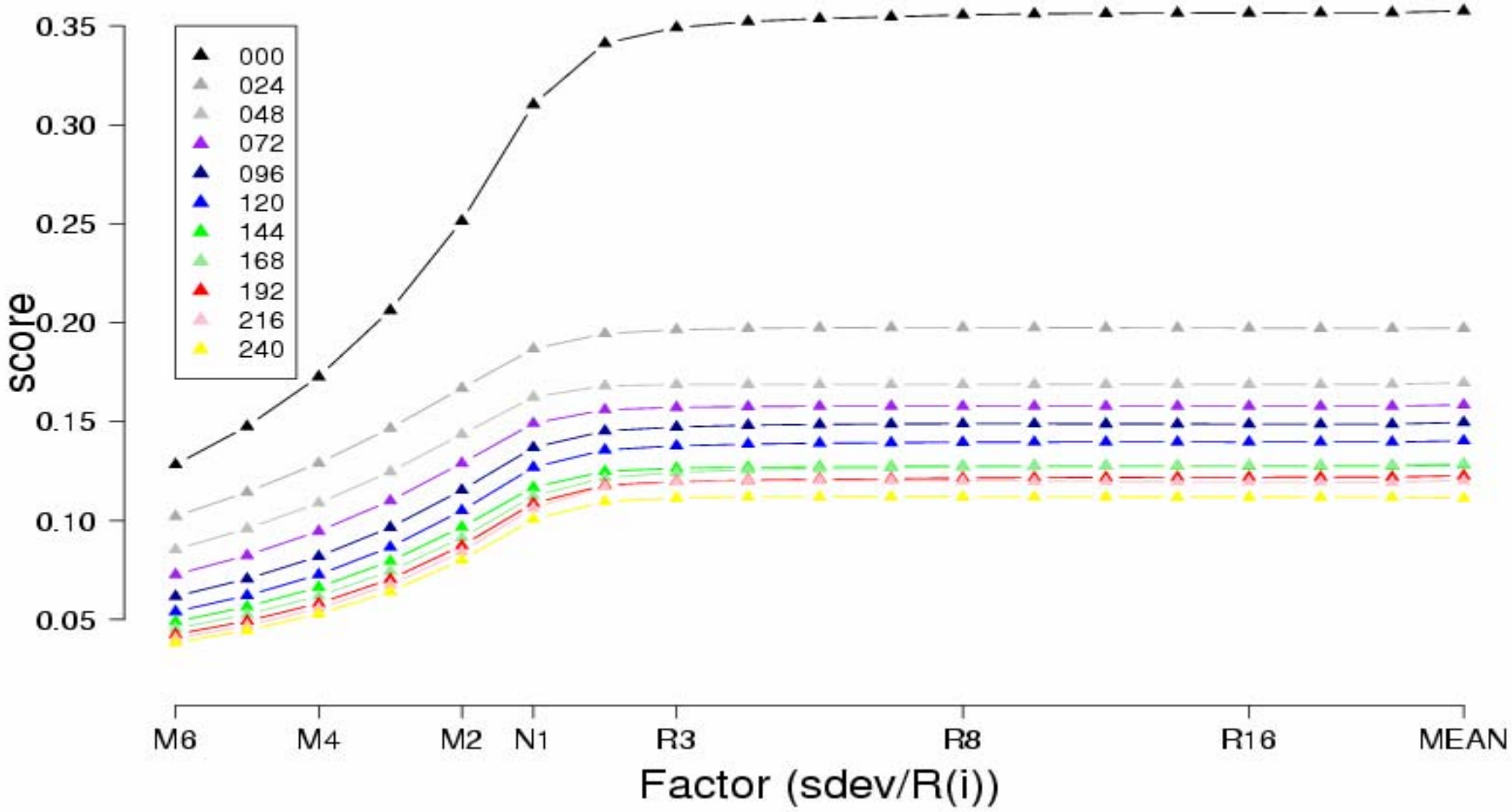
IGN = $-\log_2(P(T=T_o+/- 1))$

Canada

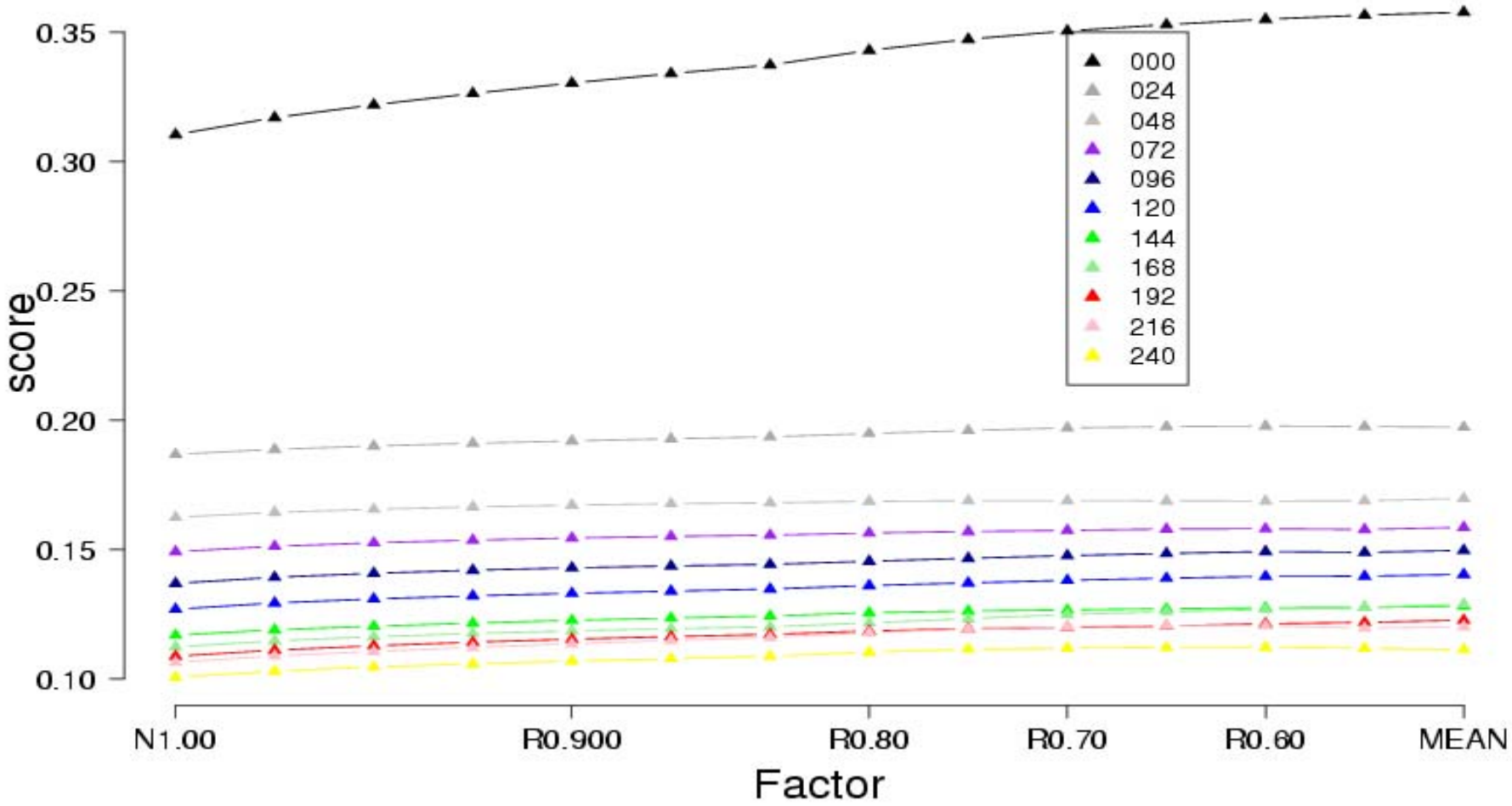
Proper Experiment (2)

- Data:
  - 90 days temperature ensemble forecasts, 16 members; 209 Canadian stations (~18000 cases)
- Assumed normal distribution
- Two methods:
  - multiplied sd by factors up to 5 and divided by up to 20 + ensemble mean
  - Truncated tails of normal distribution, added to central part.

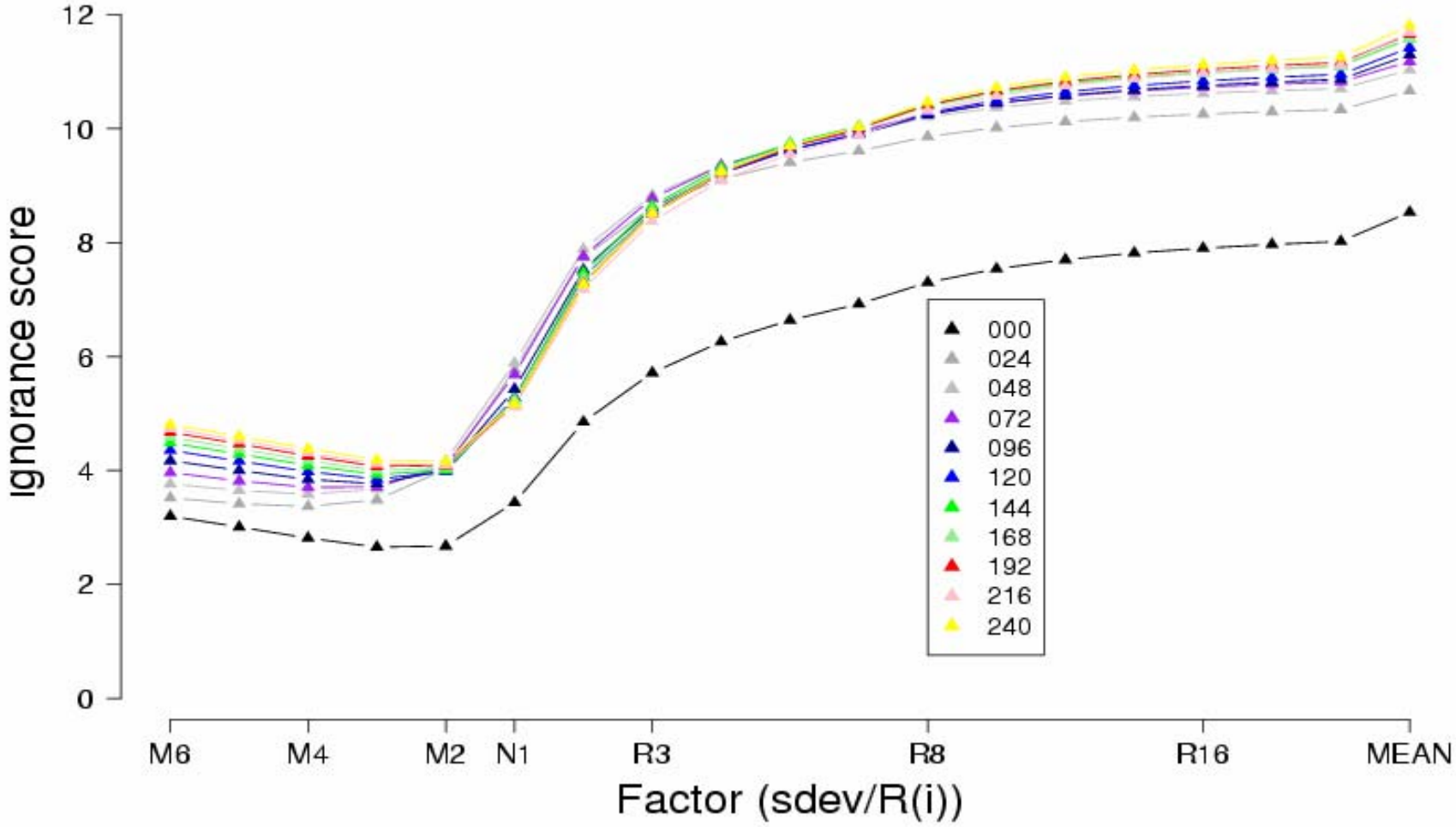
Proper experiment (3) - PS



Proper experiment (4)



Proper experiment (5)



Discussion - Is properness important?

- An alternative view:
  - Is the design of a proper score needed to offset the advantage of using prior knowledge about the variable being predicted?
  - Importance of properness vs. convenience and user-understanding of score.
  - Use of “nearly proper” scores.
- Next step
  - CRPS on same sample.

Thank you!