# Verification Basics and Issues

Bob Livezey

DOH/RDM Science Workshop
June 10, 2004
Silver Spring, MD

# Outline

- Introduction
  - Why do we do verification?
  - Forecast types
  - Performance vs. skill vs. value
  - Reference sources
- Forecast Issues
  - Quantification
  - Authentication
- Verification Issues
  - Comparison
  - Diagnosis and decomposition
  - Stratification
  - Estimation

# Why do forecast verification?

- For management purposes.

- For forecaster and forecast modeler feedback and improvement.

- For optimal customer utilization of forecast products.

# Types of Forecasts

- Value of a continuous variable.

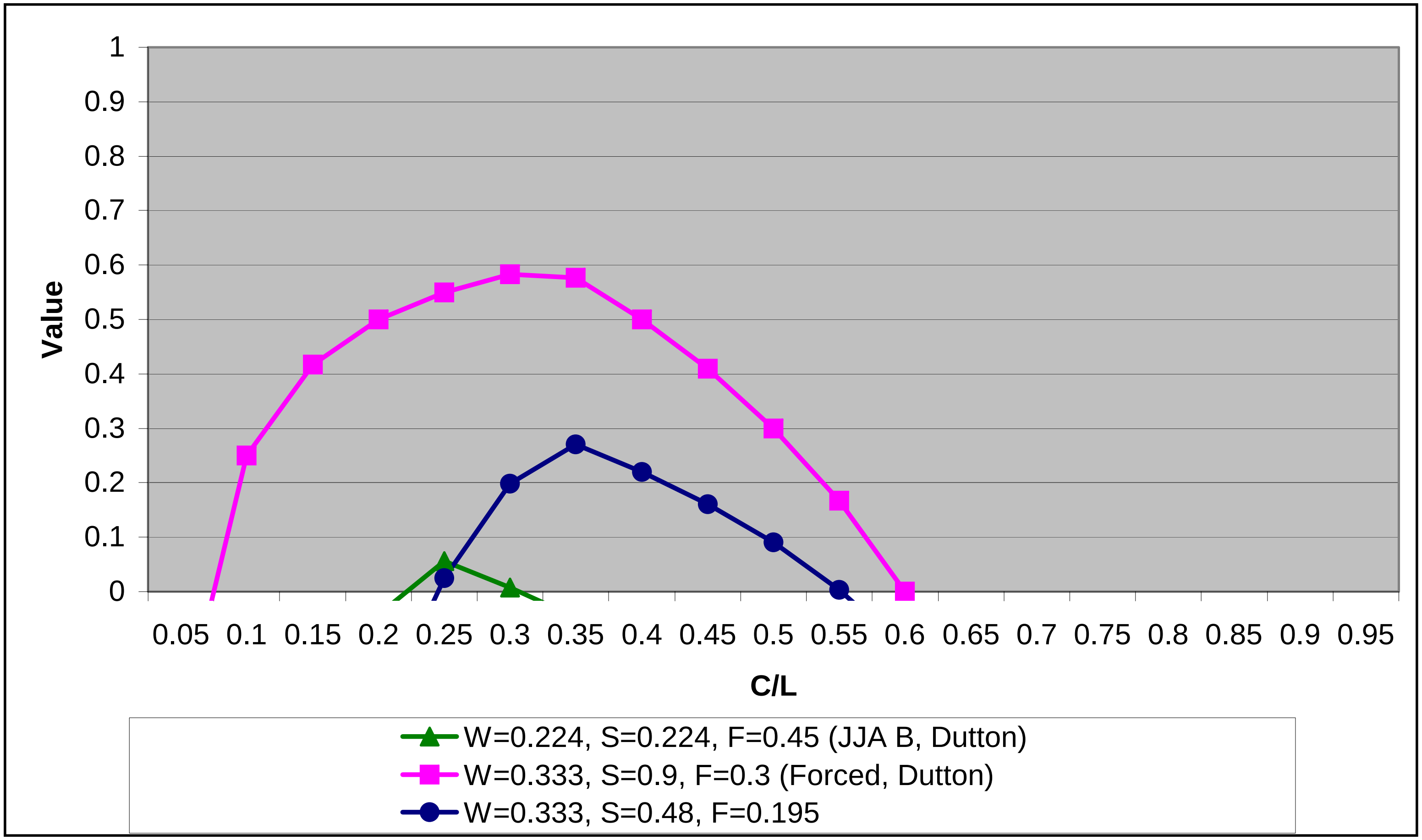
- One of two or more categories of discrete events that are mutually exclusive and collectively exhaustive.
  - Nominal (order doesn’t matter) or ordinal (order does)

- Probabilities of two or more categories of discrete events that are mutually exclusive and collectively exhaustive.

# Measures of Forecast “Goodness”

- Performance (“consistency”)
  - Average correspondence between forecasts and observations
  - Exs. Mean square error, false alarm rate, Brier score
- Skill
  - Relative performance (to either a control forecast or competitor)
  - Exs. where control forecasts are related to climatology: Mean square error skill score, Heidke skill score, rank probability skill score
- Value
  - Only has meaning in the context of a user
  - Ex. Value vs. cost/loss for binary (adverse event, no adverse event) forecasts

# Introduction to value



**W = frequency of adverse condition S = Success rate F = false alarm rate**

# Introduction to value

- Value of a forecast in a simple cost/loss environment (Dutton):

$$V = \frac{(Ec - Ef)}{(Ec - Ep)}$$

where $Ec$ = expense of climatological forecast

$Ef$ = estimated expense of actual forecast

$Ep$ = expense of perfect forecast

| | Adverse weather | |
|---|---|---|
| **Action** | **Occurs** | **Does not occur** |
| **Mitigation** | **C** | **C** |
| **None** | **L** | **0** |

# Reference Sources

- Why?
  - To optimize information return on time investment
  - To ensure use of best practices
  - To avoid reinventing verification
  - To avoid errors

- Verification
  - *Forecast Verification:  A Practitioner's Guide in Atmospheric Science*.  I. T. Jolliffe and D. B. Stephenson, Editors.  Wiley.
  - Livezey, R. E., 1999: The evaluation of forecasts.  *Analysis of Climate Variability: Applications of Statistical Techniques*, *Second Updated and Extended Edition,* Eds. H. von Storch and A. Navarra, Springer-Verlag, 179-186 and 191-198.
    - Note: Sec. 10.4 superceded by material presented in Chapter 4 of Jolliffe and Stephenson (2003).
    - Note: Contains discussion of Cross-Validation (see Estimation references)
  - Wilks, D. S., 1995b: Chapter 7, Forecast verification.  *Statistical Methods in the Atmospheric Sciences*, Academic Press, 233-281.
  - Lecture by Wilson at http://www.esig.ucar.edu/ams/shcourse.html

# Reference Sources

- Estimation and hypothesis testing
  - Livezey, R. E., 1999: *Field intercomparison. Analysis of Climate Variability: Applications of Statistical Techniques, Second Updated and Extended Edition*, Eds. H. von Storch and A. Navarra, Springer-Verlag, 161-178.
  - Wilks, D. S., 1995a: Chapter 5, Hypothesis testing. *Statistical Methods in the Atmospheric Sciences*, Academic Press, 114-158.
  - von Storch, H., and F. W. Zwiers, 1999a: Chapter 5, Estimation. *Statistical Analysis in Climate Research*, Cambridge University Press, 79-94.
  - von Storch, H., and F. W. Zwiers, 1999b: Chapter 6, The statistical test of a hypothesis. *Statistical Analysis in Climate Research*, Cambridge University Press, 99-128.
  - Lectures by Katz (Signifcance Testing), Livezey (Permutation and Bootstrap Procedures), and Mason (Cross-Validation) at http://www.esig.ucar.edu/ams/shcourse.html

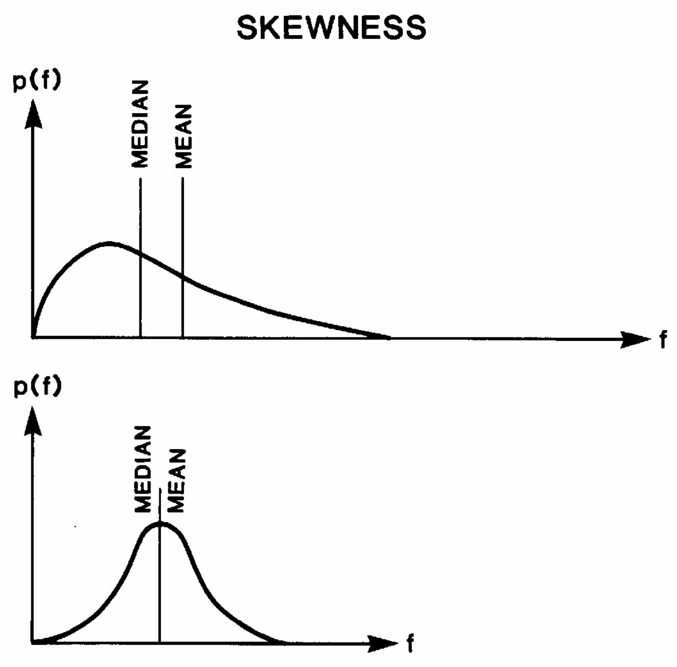
# Forecast Issues

- Can forecasts be objectively verifiable?  Are they quantifiable and unambiguous?

- Do the forecasts (or hindcasts) have any direct information about the forecast (or hindcast) period?  Are they authentic forecasts?
  - Hindcasts for statistically-based forecasts generally must be cross-validated.

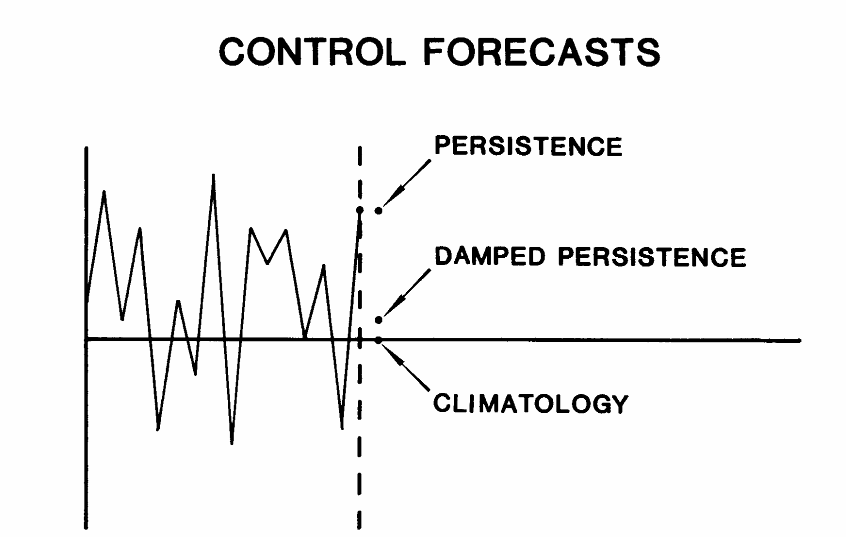
# Verification Issues -- Comparison

- Control (strawman) forecasts
  - Necessary to justify expenditure of resources, whether people time, computer time, etc.
  - Useful controls
    - Constant forecast (other than climatological normal)



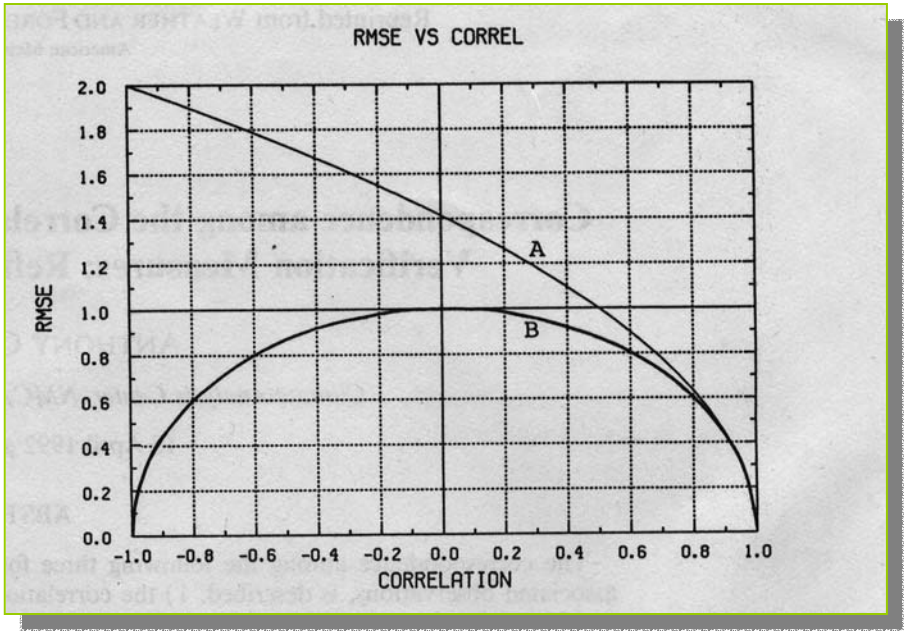
# Verification Issues -- Comparison

- Control (strawman) forecasts
  - Useful controls (continued)
    - Climatology (normal for cont. variable, random draw from distribution for categories, and distribution for probabilities)
    - Persistence
      - Anomaly persistence
      - Standardized anomaly persistence
    - Damped persistence (AR(1)/rec noise model)
    - Higher-order autoregressive models



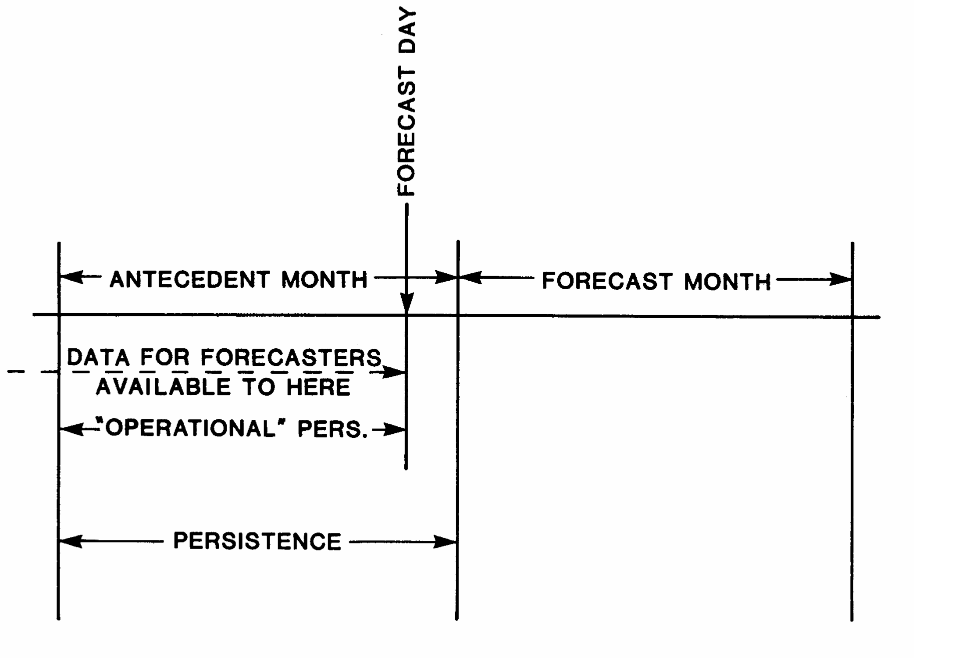
# Verification Issues -- Comparison

- Control (strawman) forecasts
  - Damped persistence generally will have smaller MSE than persistence, so it is the preferred control for MSE-based comparisons and scores.
  - Correlation must be greater than 0.5 for standardized forecasts and observations for MSE to be less than for climatology.



# Verification Issues -- Comparison

- Comparisons must be
  - Homogeneous (based on the same cases): Some cases have more predictability than others.
  - Fair: Control or competitor must have access to same information.
    - Example 1: If lagged data is used in the forecast model, a higher-order autoregressive control should be used.
    - Example 2: Operational persistence rather than persistence should be the control for old monthly forecasts.



# Verification Issues – Diagnosis and Decomposition

- Diagnostic and distribution-oriented verification
  - Forecast performance and skill are multi-faceted and should be treated as such.
  - Multiple measures and the joint distributions of forecasts and observations should be examined.

# Verification Issues – Diagnosis and Decomposition

- Example for a continuous variable forecast; decomposition of a MSE skill score

$$MSSS_j = 1 - \frac{MSE_j}{MSE_{cj}}$$

For forecasts fully cross-validated (one year at a time)

$$MSSS_j = \left\{ 2\frac{s_{fj}}{s_{xj}} r_{fxj} - \left( \frac{s_{fj}}{s_{xj}} \right)^2 - \left( \frac{\left[ \bar{f}_j - \bar{x}_j \right]}{s_{xj}} \right)^2 + \frac{2n-1}{(n-1)^2} \right\} / \left\{ 1 + \frac{2n-1}{(n-1)^2} \right\}$$

$$r_{fxj} = \frac{1}{n} \sum_{i=1}^{n} \left( f_j - \bar{f}_j \right) \left( x_j - \bar{x}_j \right) / s_{fj} s_{xj}$$

# Verification Issues – Diagnosis and Decomposition

- Example for a three-category forecast; three scores that account for increasing amounts of information applied to three different contingency tables with identical marginal distributions
  - Scores
    - CPC Heidke: Accounts only for hits and assumes climatological distribution for forecasts and observations
    - Heidke: Accounts for hits and the actual marginal distributions of the forecasts and observations
    - Gerrity: Accounts for all of the information in the contingency table

# Diagnosis & Decomposition

- Example for a three-category forecast;

| **A: Not so bad** | | Observed | | |
|---|---|---|---|---|
| Forecast | Below Normal | Near Normal | Above Normal | Forecast Dist. |
| Below Normal | 3 | 8 | 4 | 15 |
| Near Normal | 8 | 13 | 18 | 39 |
| Above Normal | 7 | 14 | 25 | 46 |
| Observed Dist. | 18 | 35 | 47 | 100 |

| **B: Bad** | | Observed | | |
|---|---|---|---|---|
| Forecast | Below Normal | Near Normal | Above Normal | Forecast Dist. |
| Below Normal | 2 | 6 | 7 | 15 |
| Near Normal | 8 | 15 | 16 | 39 |
| Above Normal | 8 | 14 | 24 | 46 |
| Observed Dist. | 18 | 35 | 47 | 100 |

| **C: Very bad** | | Observed | | |
|---|---|---|---|---|
| Forecast | Below Normal | Near Normal | Above Normal | Forecast Dist. |
| Below Normal | 0 | 6 | 9 | 15 |
| Near Normal | 8 | 15 | 16 | 39 |
| Above Normal | 10 | 14 | 22 | 46 |
| Observed Dist. | 18 | 35 | 47 | 100 |

# Diagnosis & Decomposition

- Example for a three-category forecast;

| | CPC Heidke | Heidke | Gerrity |
|---|---|---|---|
| A: Not so bad | 0.12 | 0.05 | 0.08 |
| B: Bad | 0.12 | 0.05 | 0.03 |
| C: Very bad | 0.06 | -0.02 | -0.08 |

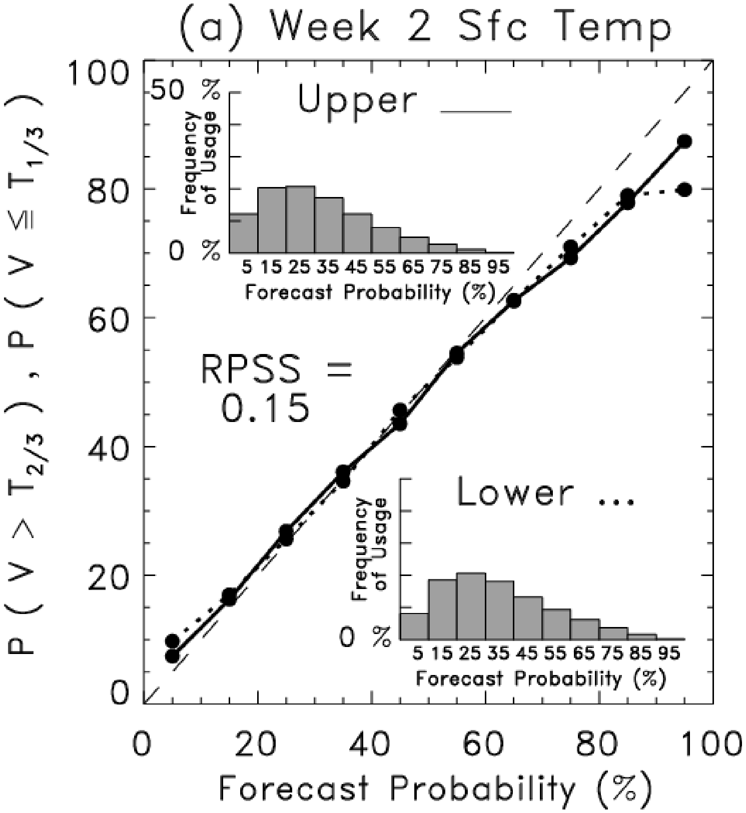
# Diagnosis & Decomposition

- Example for for probability forecasts: Calibration-refinement factorization of joint probability of forecasts and observations

$$p\left(f_i, o_j\right) = p\left(o_j \middle| f_i\right) p\left(f_i\right)$$

Calibration Refinement

(Reliability) (Sharpness)

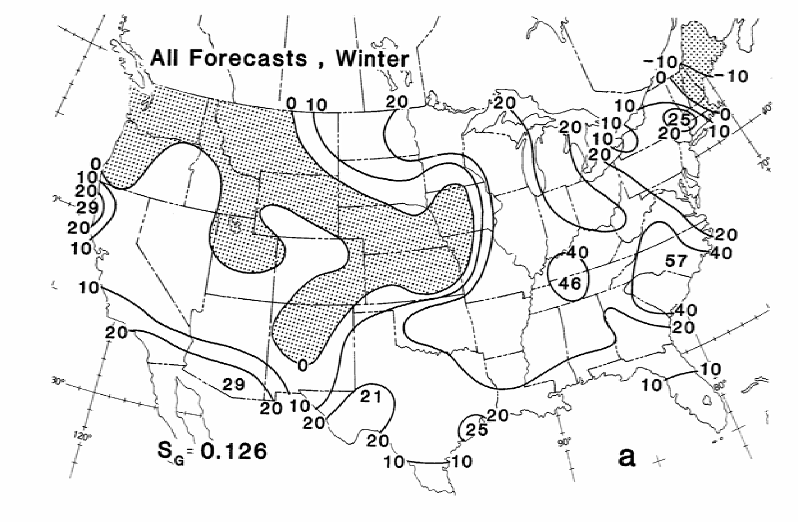


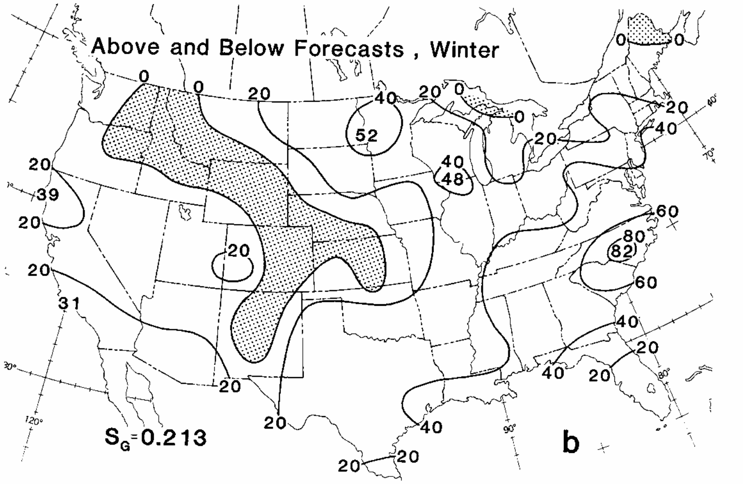
# Verification Issues – Stratification

- Important variations in performance should not be unnecessarily obscured
  - Location
  - Season
  - Situation (regime, hydro-related, etc.)

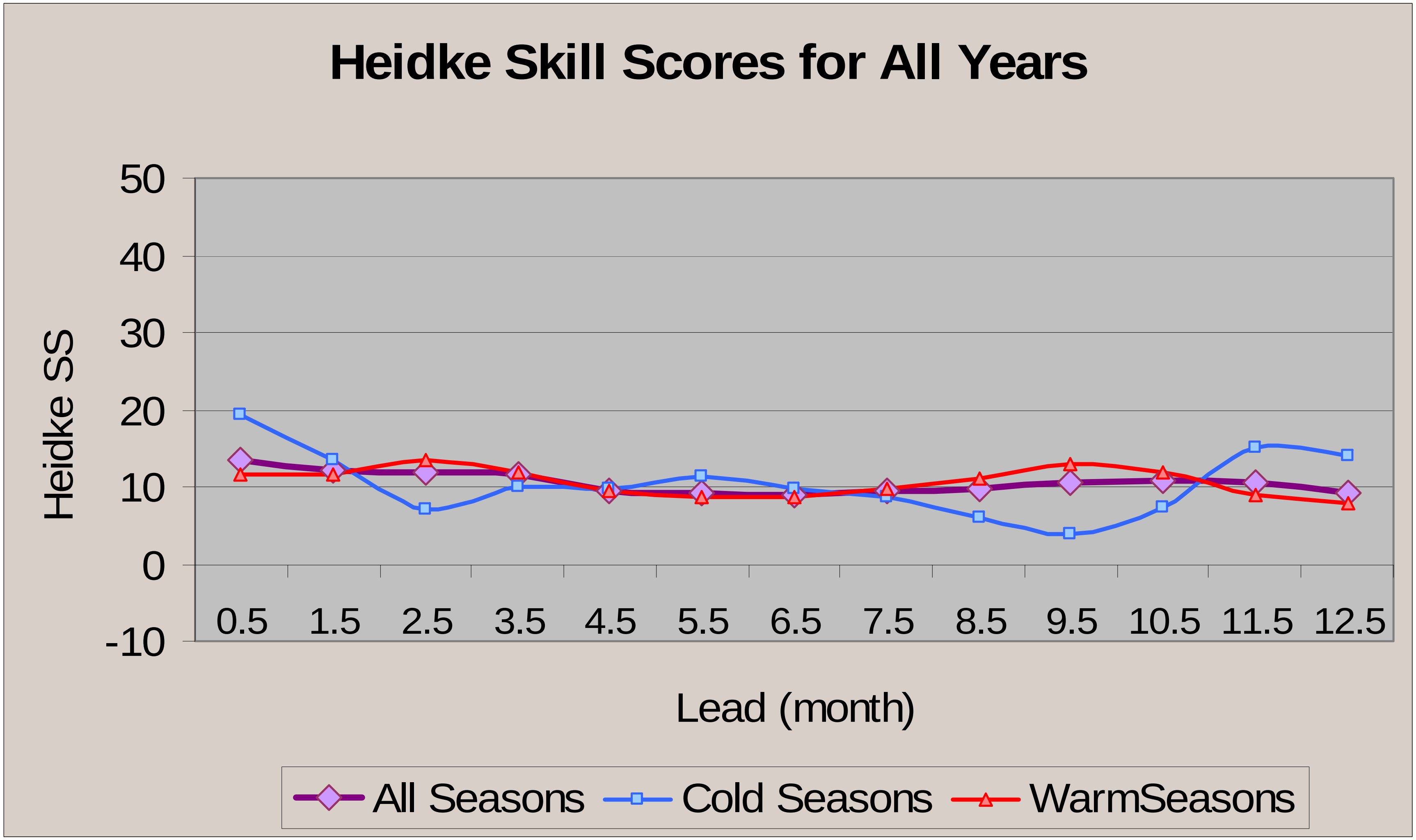
# Seasonal Temperature Forecast Skill
# 1960s to 80s

| | |
|---|---|
| All Seasons | 8.3 |
| Winter | 12.6 |
| Spring | 8.6 |
| Summer | 9.3 |
| Fall | 2.8 |

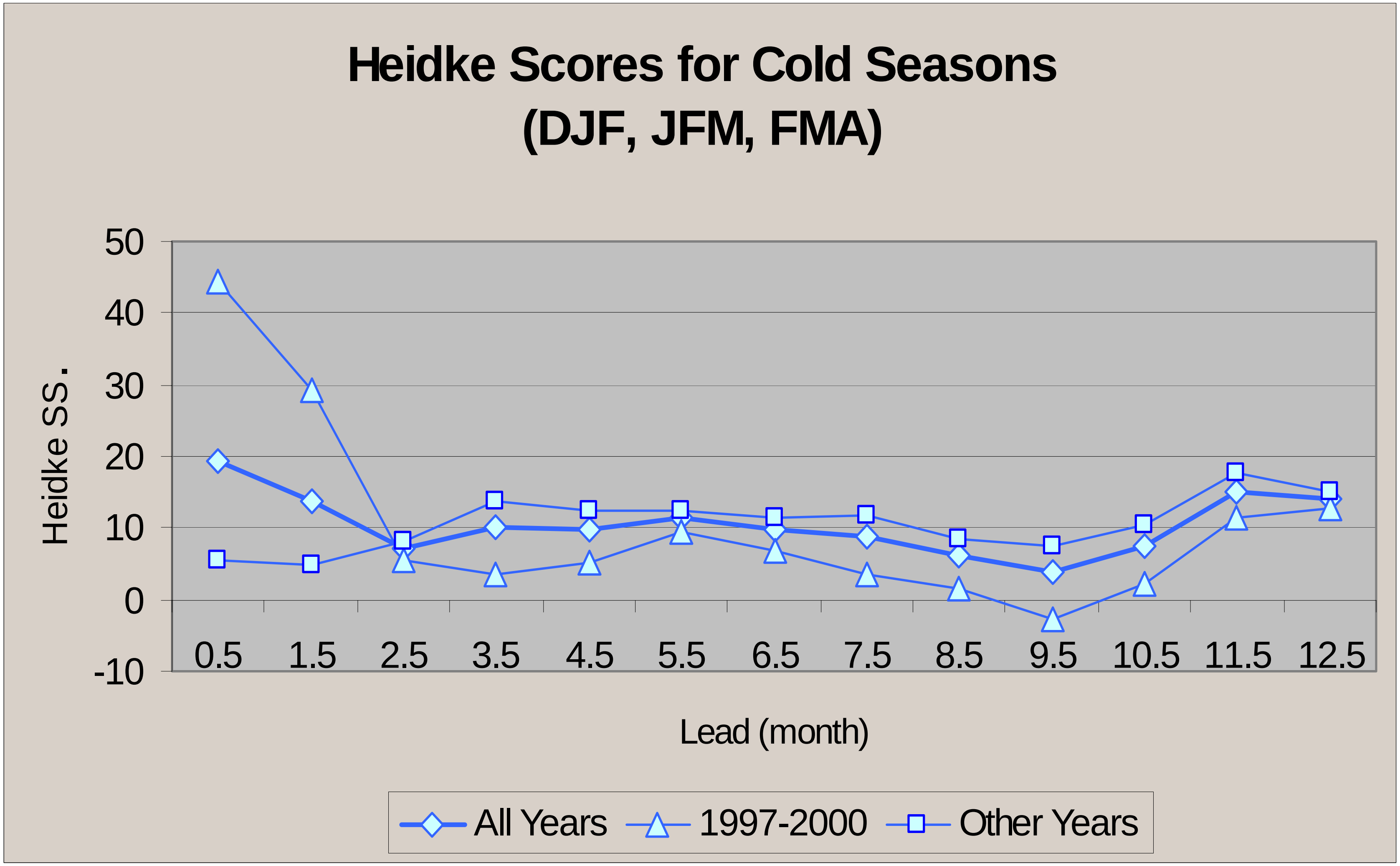




# Stratification by Lead and Seasons: Temperature



# Further Stratification by Strong-ENSO vs Other Years: Temp.

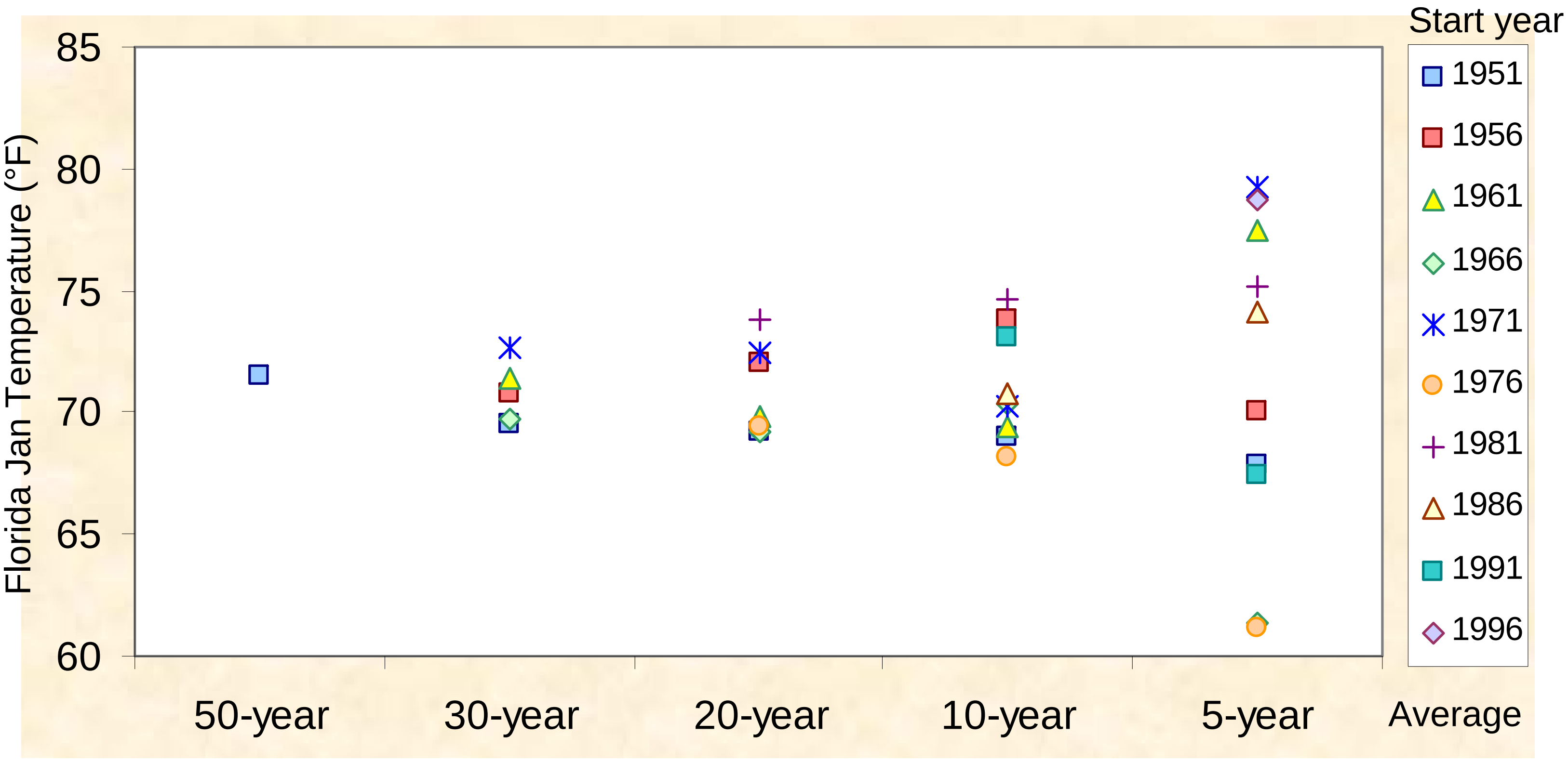


# Verification Issues – Stratification

- Stratification is inherently limited by available samples. BUT
  - reasonable tradeoffs between sample size and homogeneity are frequently possible.
  - confidence intervals can easily be estimated.

# Verification Issues – Estimation Error/Sampling Variability

- The uncertainty in an estimate of a statistic or parameter (called the *confidence interval*) increases as the sample size gets smaller and smaller:



# Verification Issues – Estimation Error/Sampling Variability

- Aggregation of data over broader and broader *time windows in the annual cycle* and over broader and broader *areas* eventually will degrade signals because of mixing climates.
- *Serial (auto-) correlation* and *spatial (cross) correlation* increase sampling uncertainty because the effective amount of independent information is less than the sample size. The sampling distribution spread is larger than for an independent sample of the same size.
- *Bootstrap procedures* are powerful and simple tools for estimating confidence intervals, including cases with serial correlation (*Moving Blocks Bootstrap Procedure).*