Statistical interpretation of NWP model output

Laurence Wilson

WMO Consultant

Outline

- What is statistical interpretation and why do it (and why not?)
- Bias removal: The easiest of all
- Linear regression: Easy and widely applicable
- Detailed example / exercise

What is Statistical Interpretation?

- Any statistical procedures that are applied to transform the output of NWP models into forecasts of weather elements that are observed at the surface or (less frequently) upper air weather elements
- Requires historical data to develop a statistical relationship
  - PREDICTAND: The element one wishes to forecast
  - PREDICTORS: Variables used to predict

Why do it?

- Fast, economical in development and application
- “Instant experience” to forecasters
- Keeps models “honest” by relating output to reality according to observations
- Includes verification of model forecasts – logical step from it

Why not do it?

•Forecasts valid only where data available

•Weaknesses forecasting extreme weather

•BUT….

•These weaknesses can be partially overcome

Requirements

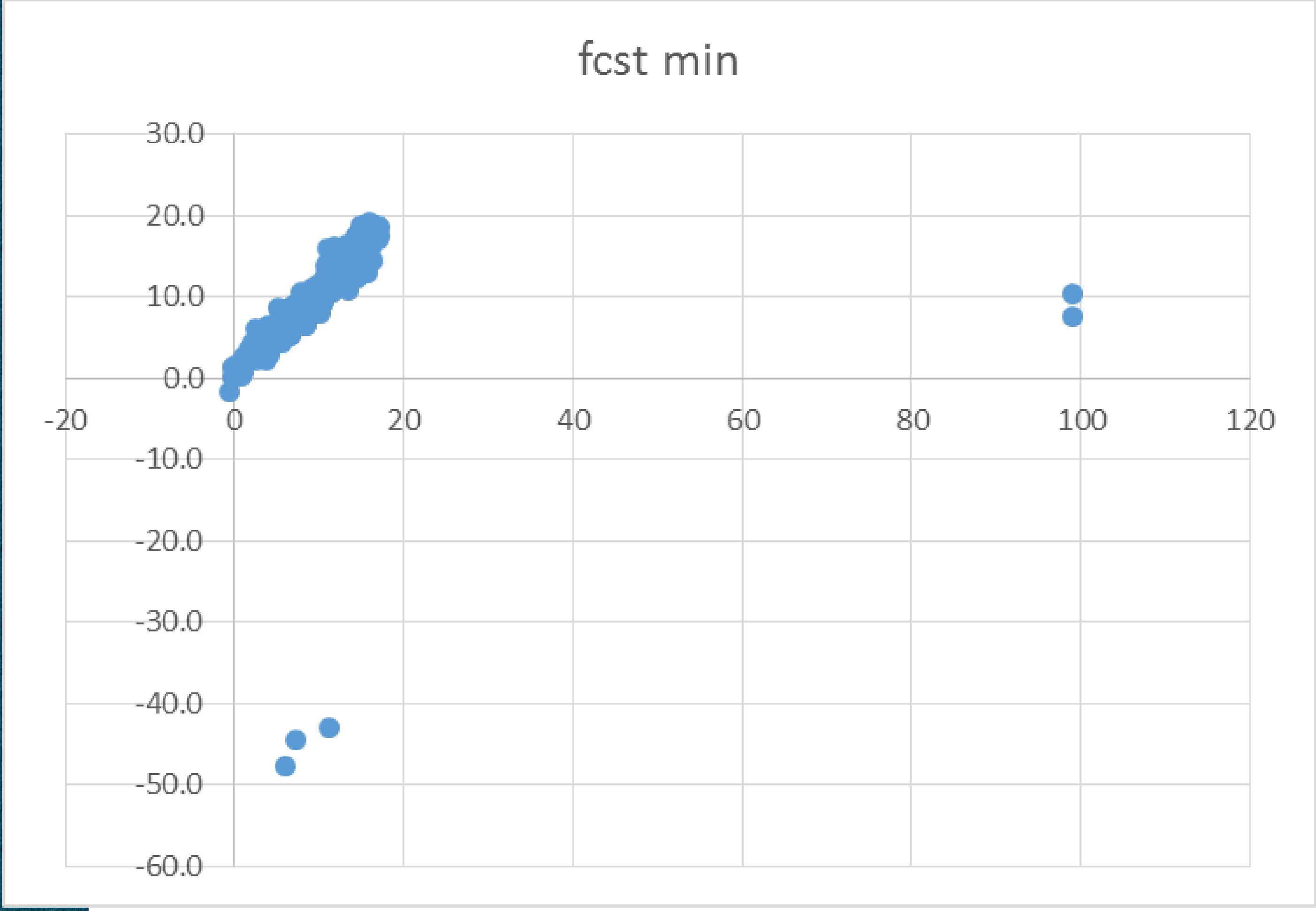
- Historical dataset, time-matched predictors and predictand
  - Similar to verification!
- statistical software (Excel can do quite a lot!)
- any not-too-old PC
- some knowledge of statistical methods and verification

STEP 1 – Get Data

- Predictand:
  - From synoptic surface observations
  - E.g. max-min temperature from station Irene
    - Decide exactly what variable to predict – max/min, wind speed, hourly “spot” temperature, highest wind speed over a period of time.
- Predictors:
  - Desired characteristics:
    - Linear relationship to predictand (e.g. model counterpart prediction – temperature, wind, precip etc.)
    - Reliably predictable – based on verification results?
    - Available (historical data for development AND operationally to run the equations)
    - Can use combination of model forecasts and observed or analysed quantities
      - As long as the analysis variable is available in predicted form.
    - …but let’s keep it simple for now – use one predictor – the model forecast
    - MOS!

STEP 1a

- MAKE sure the data is good – get rid of bad values
  - Best way – scatterplot



STEP 2 - Data preparation

- Put data into columns, predictor value and corresponding observation on the same row. (matching – just like for verification)
  - Usually a timeseries by valid time or forecast time
  - Definition must match – max over a period, point, area
  - How do we do this for Irene?
- Desirable characteristics:
  - Training sample should be representative
    - E.g. separate seasons
  - Large enough sample: ~250 cases
    - Pool data over several stations
  - As independent as possible.
- For MOS, there will be a different equation for each forecast projection – forecasts matched by valid time.

STEP 3 – Develop the equations

- Compute the bias on a representative sample and remove it.
- Simple regression is easily done in Excel

For multivariate methods (many predictors, one predictand),

- Use a package:
  - "R" probably best www.r-project.org

STEP 4 – Test the equations

- "Goodness" of fit statistics
  - "Reduction of variance" or "variance explained" or "R-squared"
- Testing on independent data
  - Essential step
  - Verification, either in real time or on independent dataset.

Statistical equations are always FIT to a training sample - They don't work as well on independent data

Bias Removal

- Bias = avg fcst – avg observed

- Estimated from the training sample
- Must be representative
  - Large and homogeneous sample
  - Seasonal stratification
  - Recent cases – dynamic bias removal
  - Weighted or not…
- Y(cor) = Y – bias
- MOS – make sure to apply the appropriate correction for the forecast projection – day 1 bias will be different from day 2 bias.

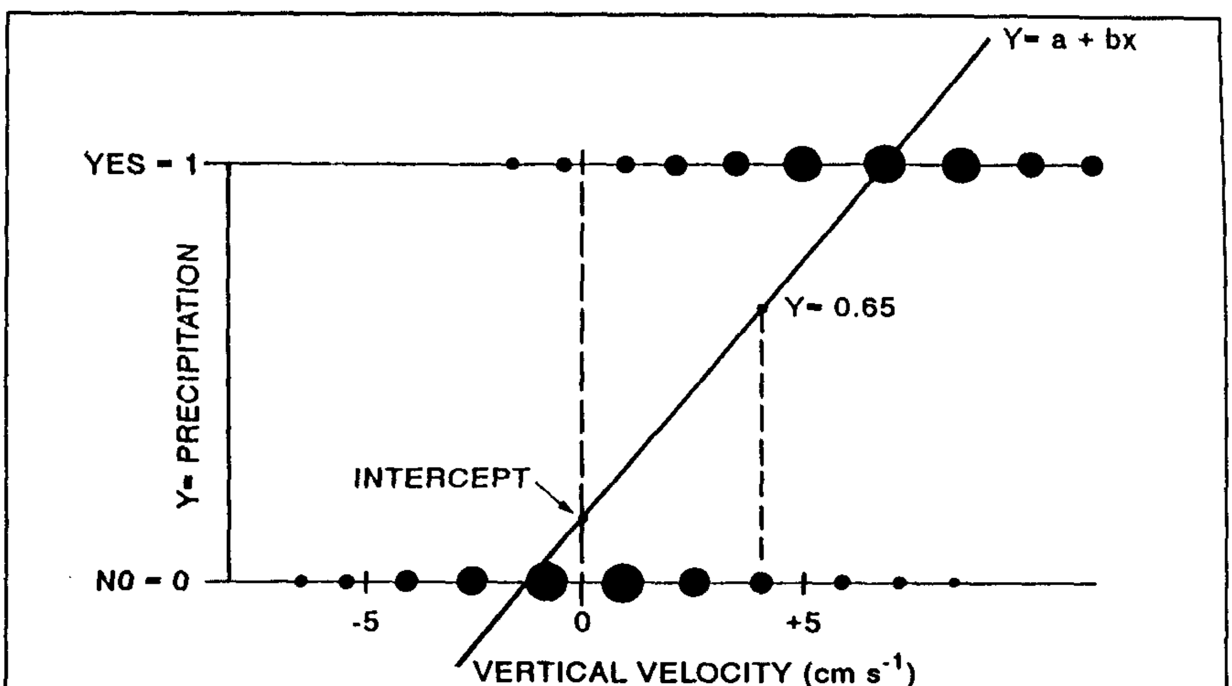
Linear Regression

- Best for continuous variables:
  - Max – min temperature
  - Temperature
  - Wind speed

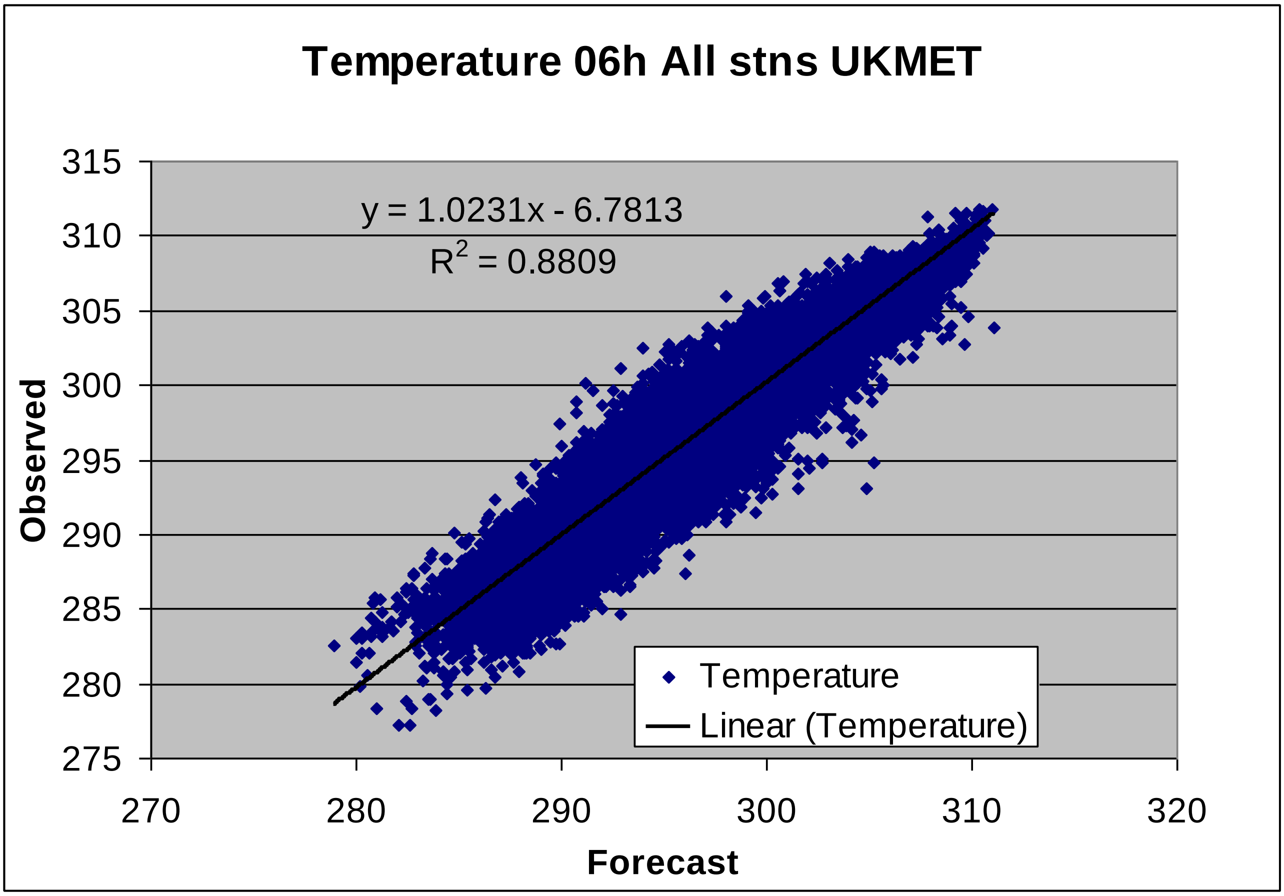
- $Y = a + bX$
  - Y is the predictand
  - a is the intercept
  - b is the slope
  - X is the predictor
- Regression automatically removes bias
- More than one predictor – multivariate regression (MLR)

Linear Regression - REEP

- REEP = Regression estimation of event probabilities
  - Predictand and maybe predictors are binary
- Example:
- Short range (33 h)
- $\hat{Y} = b_0 + b_1(\text{Thick } 1000 - 925) + b_2(T2m)$
  - RV= 0.90



SWFDP – MOS Temperature forecasts for African stations



Example and Exercise

- Max and Min temperature forecasts for station Irene (Pretoria)
- Data
  - Stn observations max/ min over 24h
  - Corresponding forecasts from UK global model, interpolated to the stn
  - Jan to Oct, 2014, 293 cases after qc
- Demo: Use of Excel to compute bias and regression forecast technique (min)
- Exercise: Interpretation (max) and REEP

Excel Template

- The template from the exercise can be used for any regression or bias correction
- To note:
  - When copying columns of data into the input data columns b and c (forecast and observed), make sure to erase any data, if your dataset is shorter than the previous one
  - The slope “LINEST” and intercept “INTERCEPT” functions require pointing to the exact row range of the data.
  - Some labels on the charts are manually inserted

Summary

- Introduction to statistical interpretation
- Simple to do in Excel
- Gives an objective correction to model errors, based on actual observations

Thank you!

Statistical Formulation Methods

Table 2. Formulation methods for statistical short range forecast equations.

	Classical	Perfect Prog	MOS
Development of Equations	Predictand- Observed at T_0 Predictors- Observed (Anal) at T_0-dT	Predictand- Observed at T_0 Predictors- Observed (Anal) at T_0	Predictand- Observed at T_0 Predictors- Forecast values valid at T_0 from prog issued at T_0-dT
Application in operat-ional fore-cast mode	Predictor values observed now(T_0) to give Forecast valid at T_0+dT	Predictor values valid for T_0+dT from prog issued now to give Forecast valid at T_0+dT	Predictor values valid for T_0+dT from prog issued now to give Forecast valid at T_0+dT
Comments	dT less than or equal to 6 hours preferable unless persistence works well. Time lag built into equations	dT can take any value for which forecast predictors are available	Same application as PPM l separate equations equations used for each dT

Statistical Formulation Methods - Characteristics

Table 3. Distinguishing characteristics of the three formulation methods.

Classical	Perfect Prog	MOS
Relationships weaken rapidly as predictor predictand time lag increases	Relationships strong because only observed data concurrent in time is used	Relationships weaken with increasing projection time due to increasing model error variance
Model-independent	Model-independent	Model-dependent
does not use model output	does not account for model bias - model errors decrease accuracy	accounts for model bias
Large development sample possible	Large development sample possible	Generally small development samples depends on frequency of model changes
Access to observed or analysed variables	Access to observed or analysed variables	Access to model output variables that may not be observed