WP_RAQ_4.1

Forecast Evaluation

Paul Agnew

Overview

‘Define common skill scores for air quality forecasts and tools for evaluating high resolution forecasts’

Deliverables

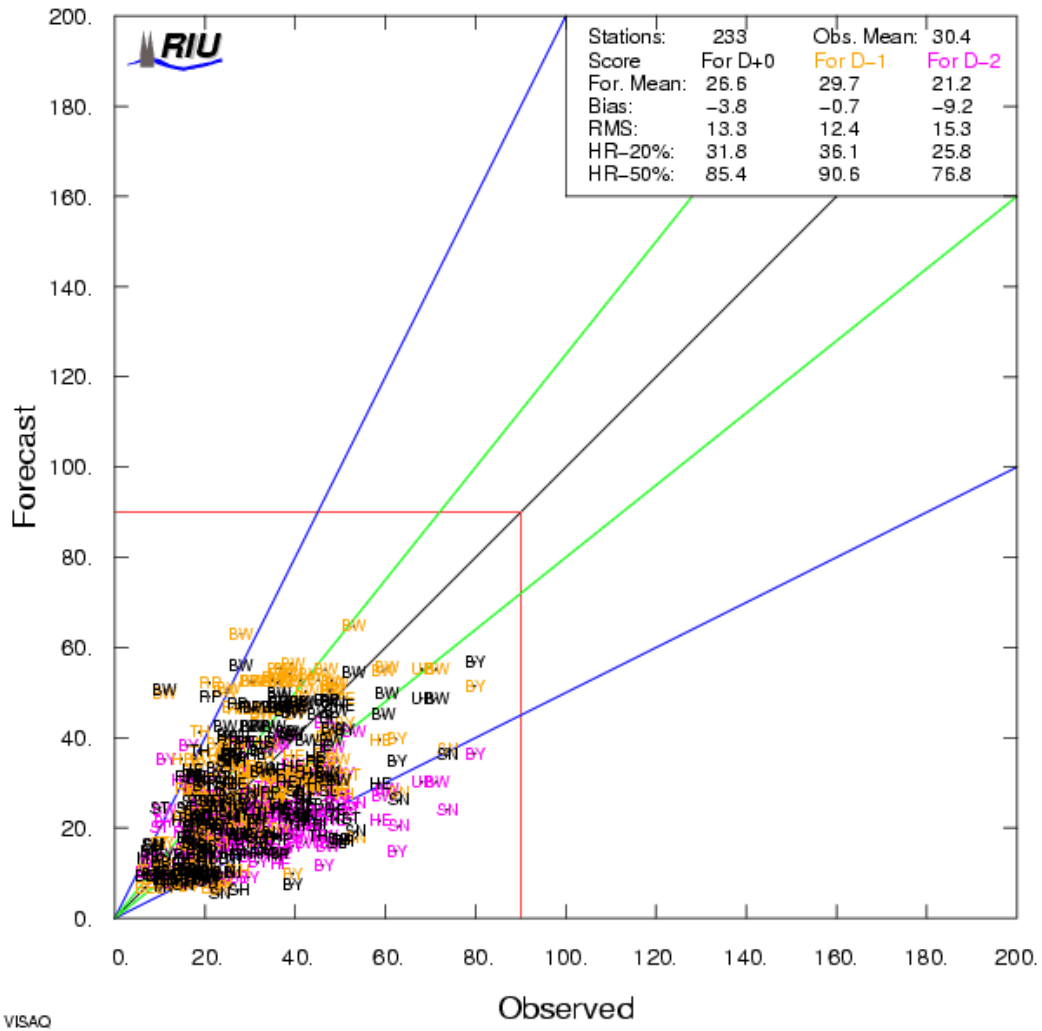
- Report on skill score characterisation for RAQ fore/hind casts
- Skill score software to compare model output and surface observations

Approach

- Utilise existing verification measures used by centres operating operational forecasts
- Literature review of alternative methodologies
- Selection of performance metrics for
  - Chemical species concentrations
  - Impact on human health
  - Crop damage indices
- Recommendations in report

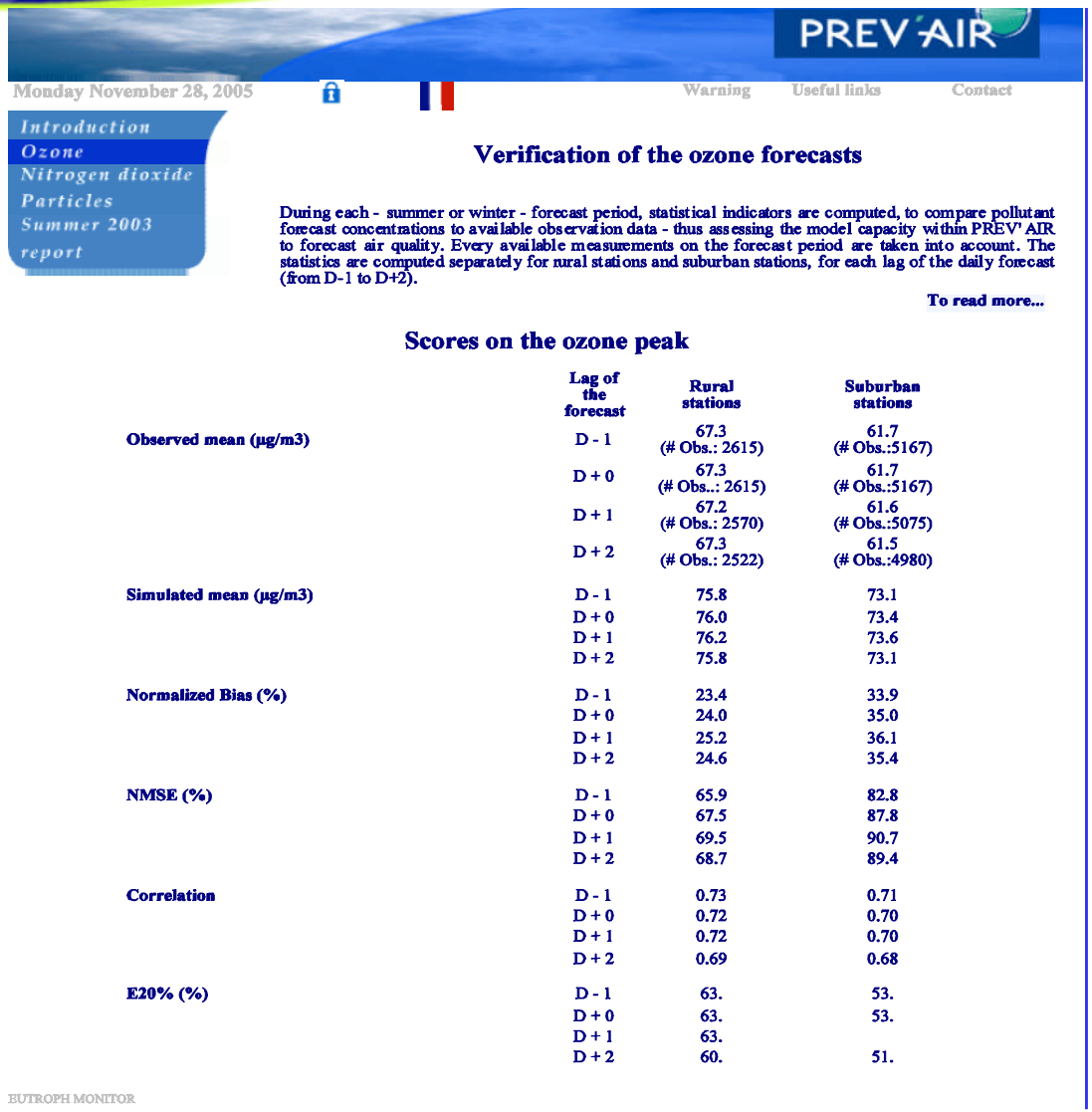
Example of existing verification - EURAD

Ozon µg/m³ Verify 27.11.2005 Daily Mean



VISAQ

Example of existing verification – Prev'Air



PREV'AIR

Monday November 28, 2005 Warning Useful links Contact

Introduction
Ozone
Nitrogen dioxide
Particles
Summer 2003 report

Verification of the ozone forecasts

During each - summer or winter - forecast period, statistical indicators are computed, to compare pollutant forecast concentrations to available observation data - thus assessing the model capacity within PREV'AIR to forecast air quality. Every available measurements on the forecast period are taken into account. The statistics are computed separately for rural stations and suburban stations, for each lag of the daily forecast (from D-1 to D+2).

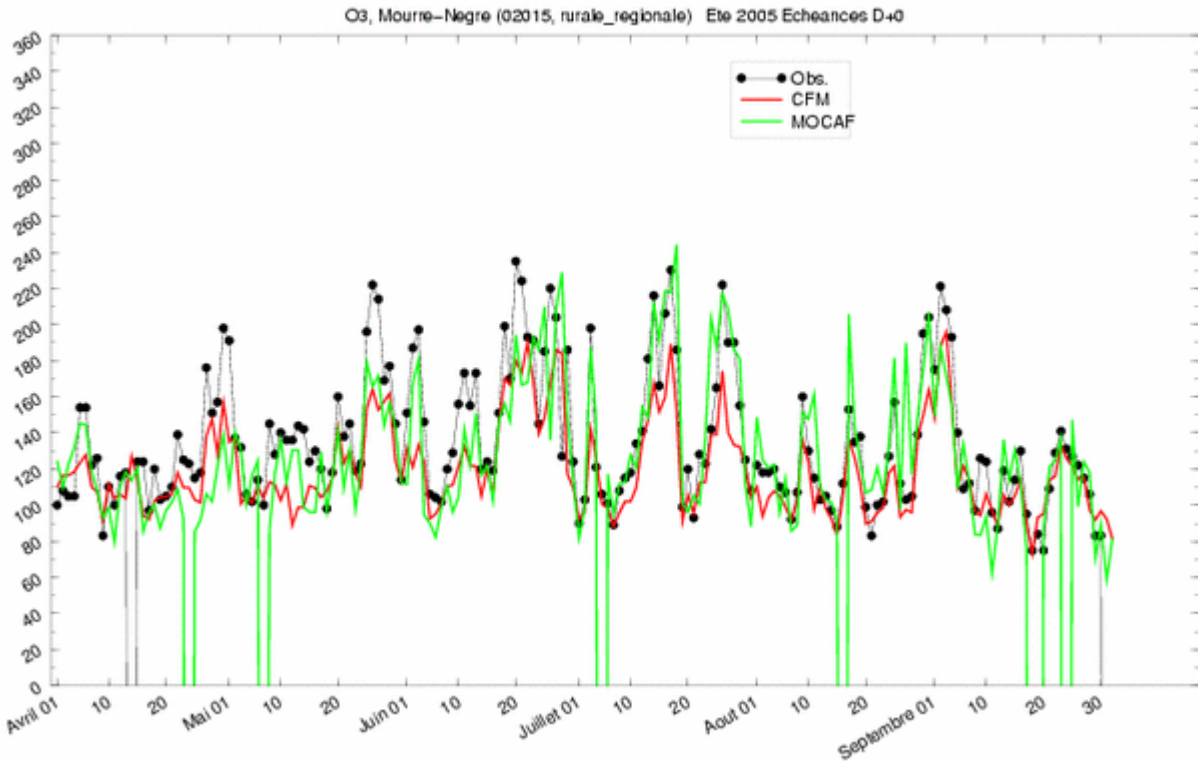
To read more...

Scores on the ozone peak

	Lag of the forecast	Rural stations	Suburban stations
Observed mean (µg/m3)	D - 1	67.3 (# Obs.: 2615)	61.7 (# Obs.:5167)
	D + 0	67.3 (# Obs..: 2615)	61.7 (# Obs.:5167)
	D + 1	67.2 (# Obs.: 2570)	61.6 (# Obs.:5075)
	D + 2	67.3 (# Obs.: 2522)	61.5 (# Obs.:4980)
Simulated mean (µg/m3)	D - 1	75.8	73.1
	D + 0	76.0	73.4
	D + 1	76.2	73.6
	D + 2	75.8	73.1
Normalized Bias (%)	D - 1	23.4	33.9
	D + 0	24.0	35.0
	D + 1	25.2	36.1
	D + 2	24.6	35.4
NMSE (%)	D - 1	65.9	82.8
	D + 0	67.5	87.8
	D + 1	69.5	90.7
	D + 2	68.7	89.4
Correlation	D - 1	0.73	0.71
	D + 0	0.72	0.70
	D + 1	0.72	0.70
	D + 2	0.69	0.68
E20% (%)	D - 1	63.	53.
	D + 0	63.	53.
	D + 1	63.	
	D + 2	60.	51.

EUTROPH MONITOR

Example of existing verification – Prev'Air time series



Verification Issues

Requirements

- Routine evaluation of forecasts c.f. observations
  - (N)RMS error, bias and correlation take into account all forecasts and observations, across the range of values
  - Sensitive to model resolution: ‘smoother’ models will have better scores overall but may under-forecast exceedance events
- Skill scores focussed on threshold exceedance events

Proposal to maintain these fundamental measures

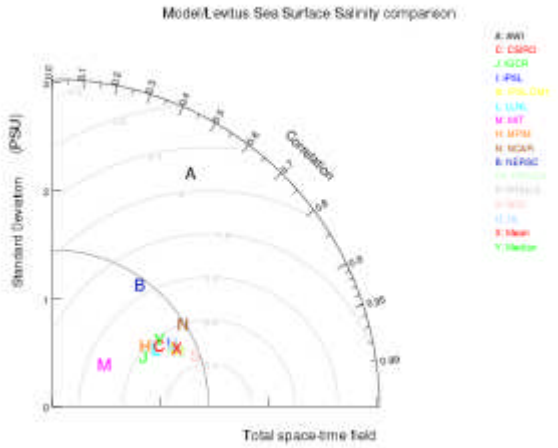
- Normalised RMSE
- Bias
- Correlation

- These fundamental verification statistics present an important summary of model performance

- How best to display this information?

Taylor Diagrams

- Summarises basic verification statistics, comparing forecast to reference fields
  - Correlation
  - Pattern NRMSE
- Use to compare a number of different models
  - Easy visual interpretation



Skill Scores

- Requirement: a single statistic indicating the relative skill of each model in forecasting threshold exceedences

- Basis: 2x2 contingency table
  - a – Hit
  - b – False alarm
  - c – Miss
  - d – Correct rejection
  - n=a+b+c+d total no. events

		Events	**Observed**
		Yes	No
Events	Yes	a	b
Forecast	No	c	d

Range of Skill Scores

- A range of indicators traditionally developed for meteorological forecasts:
  - Proportion Correct, Heidke Skill Score, Gilbert SS, Peirce (Kuipers) SS etc.
- Require a Skill Score which is:
  - Simple to calculate and interpret
  - Not sensitive to the thresholds chosen
  - Not sensitive to the ‘base rate’
  - Robust – not easily ‘hedged’
  - Can be tested for significance if required
- The ‘Odds Ratio’ meets these requirements

Odds Ratio

- ‘Odds’ defined as

  ratio of probability that event occurs to probability that event does not occur

- Odds Ratio: forecast skill can be judged by comparing odds of good forecast (hit) to odds of bad forecast (false alarm)

- Easily calculated from contingency table

- Depends solely on the conditional joint probabilities: independent of any bias between observations and forecasts

Odds Ratio Skill Score

- A skill score can be derived by a simple transformation:

  ORSS=(OR-1)/(OR+1)

  - This mapping produces a skill score in the range -1 to +1

- When ORSS=-1 forecasts and observations are independent

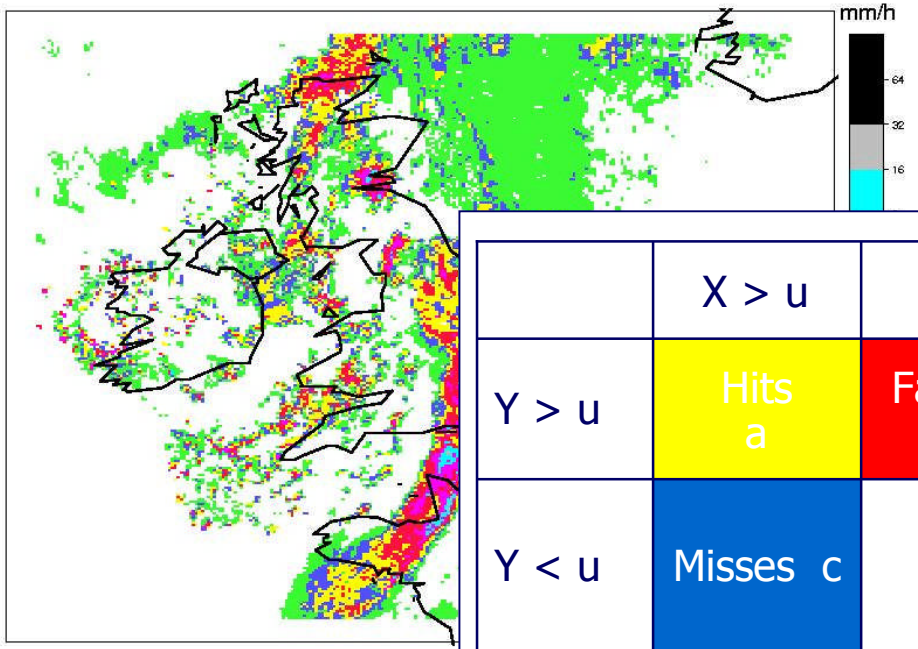
- *Providing number of forecasts is statistically significant*, ORSS approaching +1 indicates a skillful forecast

Consideration of Scale Decomposition

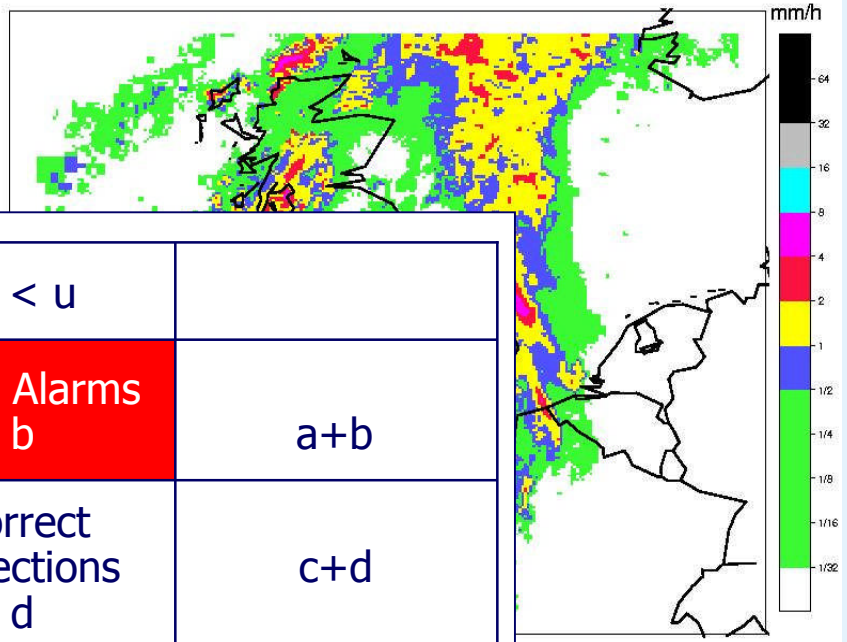
- Valuable to probe to the differing levels of skill in models at different scales

- Invoke methods of scale decomposition: increasingly used in diagnosing precipitation forecast performance

Radar

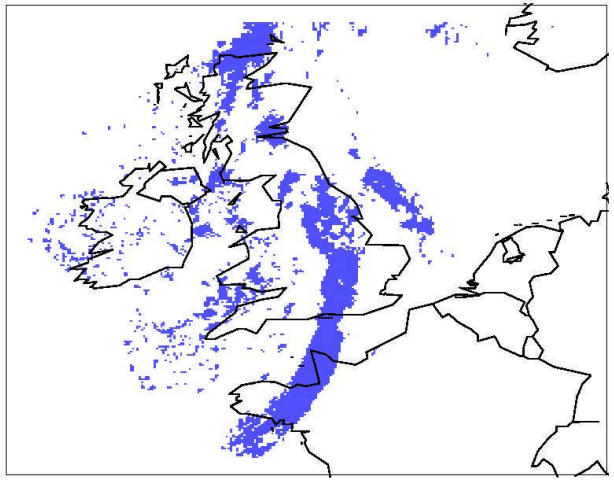


Model forecast

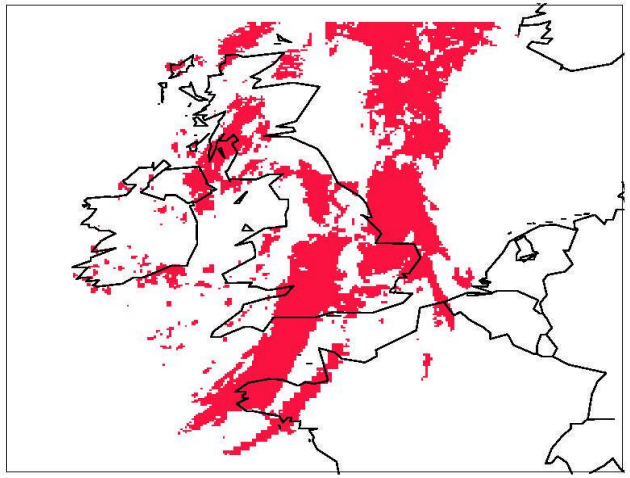


	X > u	X < u	
Y > u	Hits a	False Alarms b	a+b
Y < u	Misses c	Correct Rejections d	c+d
	a+c	b+d	a+b+c+d=n

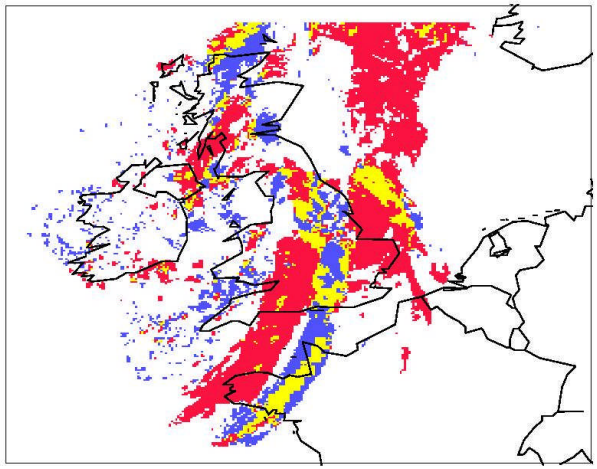
Radar > 1 mm



Forecast > 1 mm

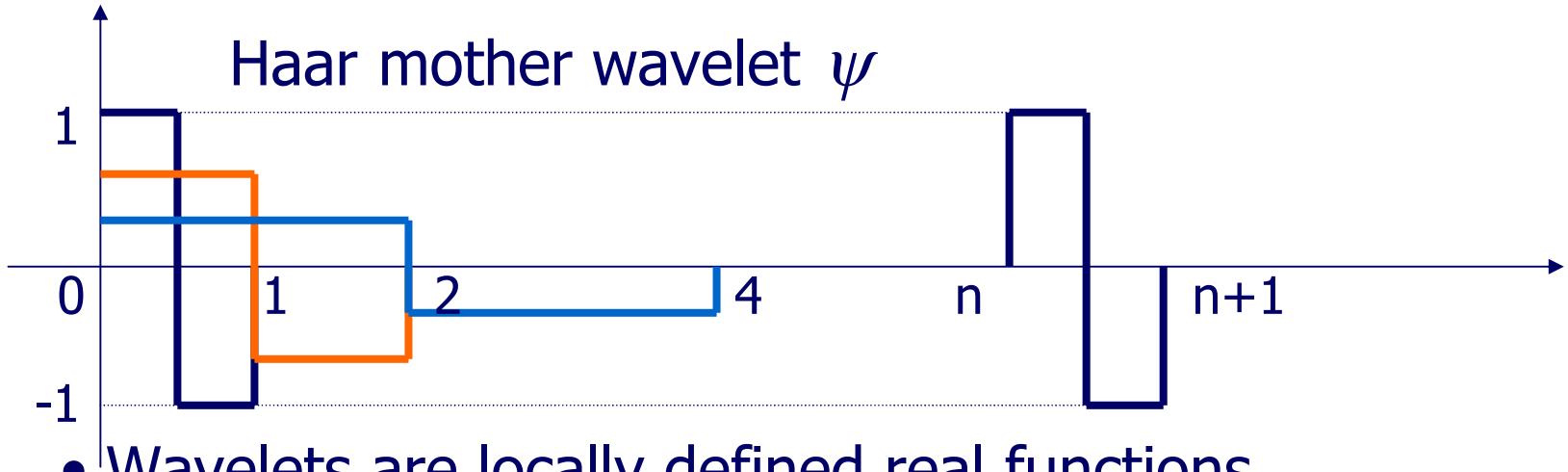


Binary error image



Source: Marion Mittermaier, derived from Casati (2004)

An intensity-scale technique using wavelets (Marion Mittermaier – Met Office 2005)



- Wavelets are locally defined real functions characterised by a **location** and a **spatial scale**.
- Any real function can be expressed as a linear combination of wavelets, i.e. as a sum of components with different spatial scales.
- Wavelet transforms deal with discontinuities better than Fourier transforms do

Scale Decomposition

- Technique is valuable as a *detailed* diagnostic for probing the scale at which models exhibit/fail to exhibit skill
- Requires field to verify against (in precip. typically provided by radar imagery)
- Not yet a sufficiently mature methodology for use as a routine indicator of comparative forecast skill

Proposed Verification: Basic chemical species

- For O_3, SO_2, NO_2, PM10, CO
- Verify against station data: forecast field data interpolated to station point
- Stratification by
  - Lead time (24,48,72 Hour)
  - Type of site (urban vs rural)
- Taylor Diagrams to summarise verification of daily fields (00Z and 12Z)
- NRMSE, Bias, Correlation time series for each partner model – assess on-going performance
- Baseline comparison: 24 hour persistence forecast

Proposed Verification: Skill score

- Odds Ratio Skill Score based on contingency table for forecast/observed exceedence of information and warning threshold at observation sites
  - Which species? All species?
- Sum individual ORSS over all observation sites and normalise
- Display time series for each partner model

Implementation: Verification Software

- Core verification performed centrally at ECMWF
- New tools developed using ‘MetPy’
  - User-friendly scripting language
  - Full functionality via numerical/statistical libraries
  - Straight-forward publishing of verification measures on GEMS RAQ web pages
- Potential for partners to develop tailored verification measures, running MetPy on ecgate – interest?

Language

```
compute(
        param = Z,
        levtype = pl,
        levelist = (1000,500,100),
        score = (ancf,ref),
        steps = StepSequence(12,240,12),
        area = (‘europe’, ‘north hemisphere’),
        forecast = forecast (
        )
        persistence = persistence(
        )
        analysis = analysis (
                expver = ‘0001’,
                date = DateSequence(20040101,20040131),
        )
)
```

Verification Implementation: Requirement

- Technical specification document
  - Summarising required verification metrics
  - Stratification of data
  - Structure of web pages

Report Outline

- Introduction
- Review existing procedures (incl. questionnaires)
- Results of literature review
- Review of impact metrics
  - Human health
  - Crop damage
- Issues related to observation sites
- City level forecast issues
- Recommendations