Evaluation of DELTA Forecasting MQO v5.5

forecasting system evaluation project challenges

Jenny Stocker,
Kate Johnson &
Amy Stidworthy

FAIRMODE Technical Meeting

June 2017

Athens

Greece

Contents

- Context
- Threshold criteria
- System evaluation
- Flexibility options
- ‘To be discussed at meeting’
- Summary

Context

- Many improvements have been implemented in the forecasting mode of the DELTA Tool i.e. it is now more robust in terms of what it calculates
- **How suitable is it for use in evaluating a forecasting system**?

- CERC undertook a project to perform an 'Evaluation of point-wise Air Quality Index for Health forecast data'
- **Project for the *Irish Environmental Protection Agency* (Kevin Delaney, Patrick Kenny)**
- Forecast ozone, NO_2, PM_{10}, $PM_{2.5}$ and SO_2 at 12 sites in Ireland
- Contracted to use both the DELTA Tool and the Model Evaluation Toolkit*
- The project highlighted the positive and negative aspects of both tools

- In January 2017, CERC worked with Stijn & Philippe on the outstanding issues with the tool:
  - Some have been resolved in DELTA Tool version 5.5
  - Some items remain open

airTEXT

CERC

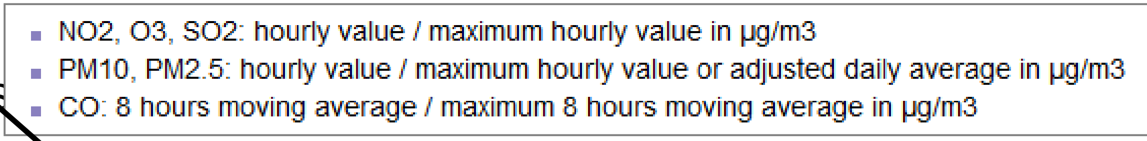
* Freely downloadable from www.cerc.co.uk/ModelEvaluationToolkit

Threshold criteria

- What are we evaluating against i.e. what are our threshold criteria?
- These differ across Europe:
  - Threshold names
  - Threshold values
  - Index values
  - Pollutant averaging times

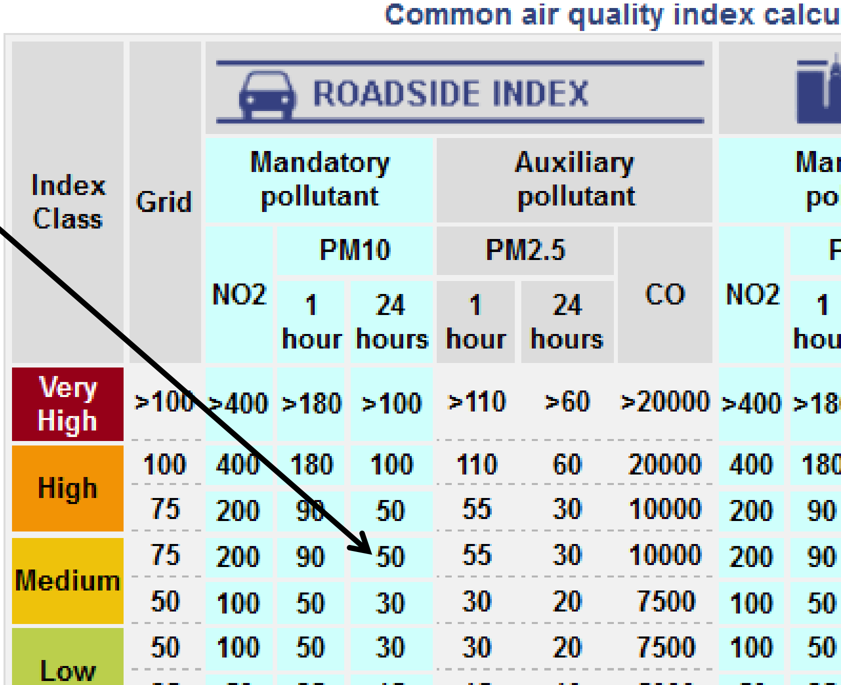
Common Air Quality Index (CAQI) (2006)

THE HOURLY AND DAILY COMMON INDICES



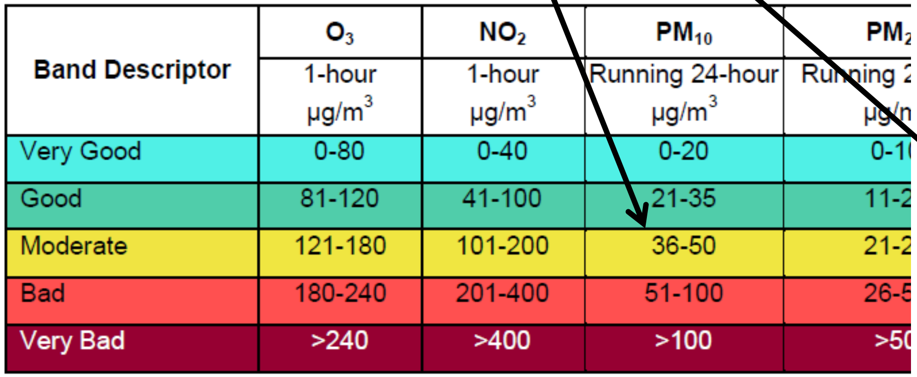
- NO2, O3, SO2: hourly value / maximum hourly value in µg/m3
- PM10, PM2.5: hourly value / maximum hourly value or adjusted daily average in µg/m3
- CO: 8 hours moving average / maximum 8 hours moving average in µg/m3

Common air quality index calcu



Index Class	Grid	ROADSIDE INDEX							
		Mandatory pollutant			Auxiliary pollutant			Mar po	
		NO2	PM10		PM2.5		CO	NO2	P
			1 hour	24 hours	1 hour	24 hours			1 hou
Very High	>100	>400	>180	>100	>110	>60	>20000	>400	>18
High	100	400	180	100	110	60	20000	400	180
	75	200	90	50	55	30	10000	200	90
Medium	75	200	90	50	55	30	10000	200	90
	50	100	50	30	30	20	7500	100	50
Low	50	100	50	30	30	20	7500	100	50

Table 2-1 AQI component pollutants, bands and colours used in the prototype.



Band Descriptor	O_3	NO_2	PM_{10}	PM_2
	1-hour µg/m^3	1-hour µg/m^3	Running 24-hour µg/m^3	Running 2 µg/n
Very Good	0-80	0-40	0-20	0-1
Good	81-120	41-100	21-35	11-2
Moderate	121-180	101-200	36-50	21-2
Bad	180-240	201-400	51-100	26-5
Very Bad	>240	>400	>100	>50

Prototype EU Air Quality Index (2016)

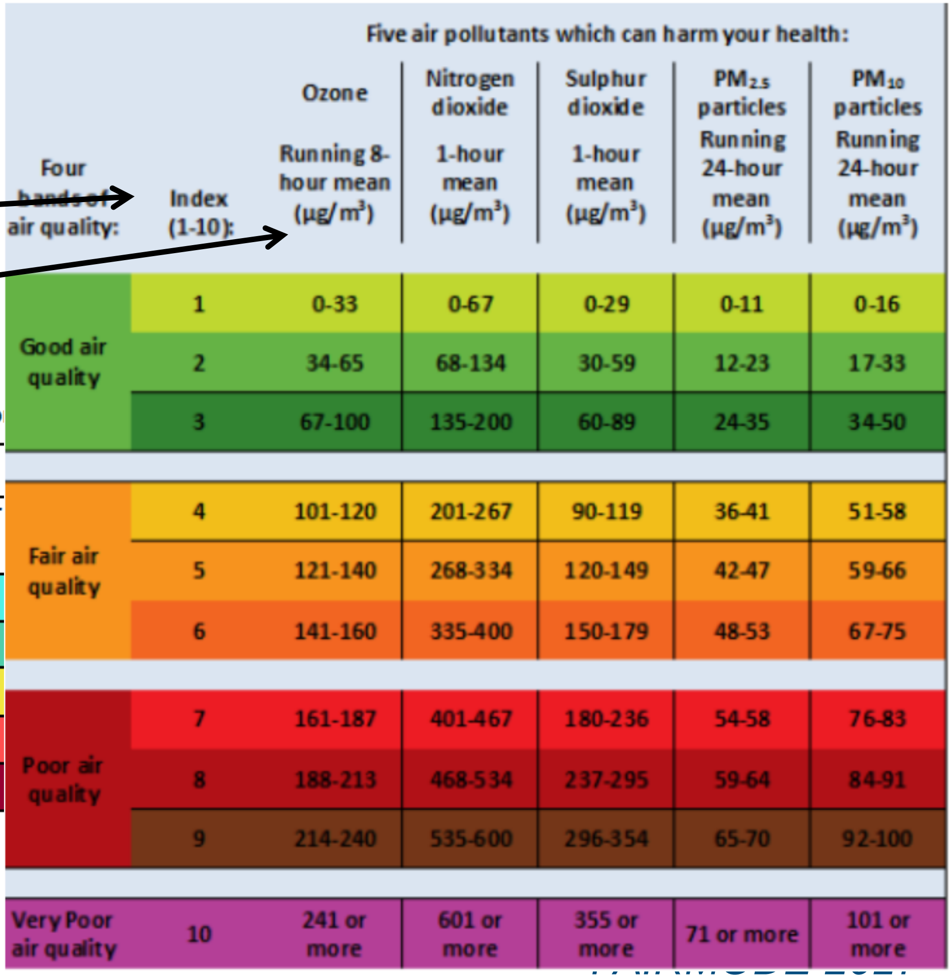
(Ricardo report for DG ENV)

CERC

Threshold criteria

- What are we evaluating against i.e. what are our threshold criteria?
- These differ across Europe:
  - Threshold names
  - Threshold values
  - Index values
  - Pollutant averaging times

Irish Air Quality Index for Health



Five air pollutants which can harm your health:						
		Ozone	Nitrogen dioxide	Sulphur dioxide	$PM_{2.5}$ particles	PM_{10} particles
Four bands of air quality:	Index (1-10):	Running 8-hour mean (µg/m³)	1-hour mean (µg/m³)	1-hour mean (µg/m³)	Running 24-hour mean (µg/m³)	Running 24-hour mean (µg/m³)
Good air quality	1	0-33	0-67	0-29	0-11	0-16
	2	34-65	68-134	30-59	12-23	17-33
	3	67-100	135-200	60-89	24-35	34-50
Fair air quality	4	101-120	201-267	90-119	36-41	51-58
	5	121-140	268-334	120-149	42-47	59-66
	6	141-160	335-400	150-179	48-53	67-75
Poor air quality	7	161-187	401-467	180-236	54-58	76-83
	8	188-213	468-534	237-295	59-64	84-91
	9	214-240	535-600	296-354	65-70	92-100
Very Poor air quality	10	241 or more	601 or more	355 or more	71 or more	101 or more

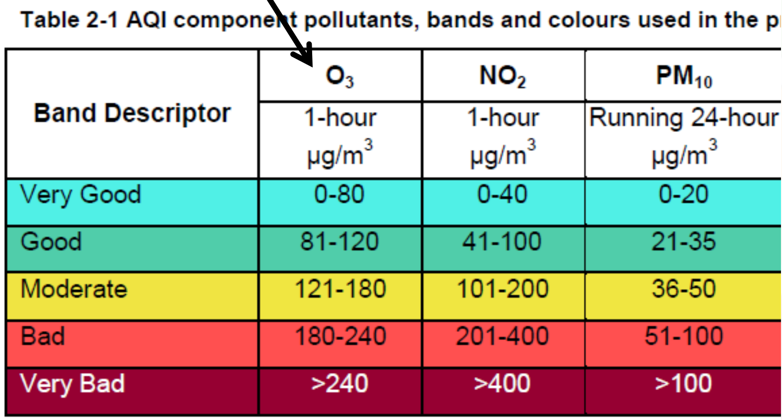


Table 2-1 AQI component pollutants, bands and colours used in the p

Band Descriptor	O_3	NO_2	PM_{10}
	1-hour µg/m³	1-hour µg/m³	Running 24-hour µg/m³
Very Good	0-80	0-40	0-20
Good	81-120	41-100	21-35
Moderate	121-180	101-200	36-50
Bad	180-240	201-400	51-100
Very Bad	>240	>400	>100

Prototype EU Air Quality Index (2016)

CERC **(Ricardo report for DG ENV)**

Threshold criteria

- What are we evaluating against i.e. what are our threshold criteria?
- These differ across Europe:
  - Threshold names
  - Threshold values
  - Index values
  - Pollutant averaging times

In the DELTA Tool:

- Each pollutant is run separately
- Each threshold is entered separately
- A lower threshold will include the higher exceedance values e.g.

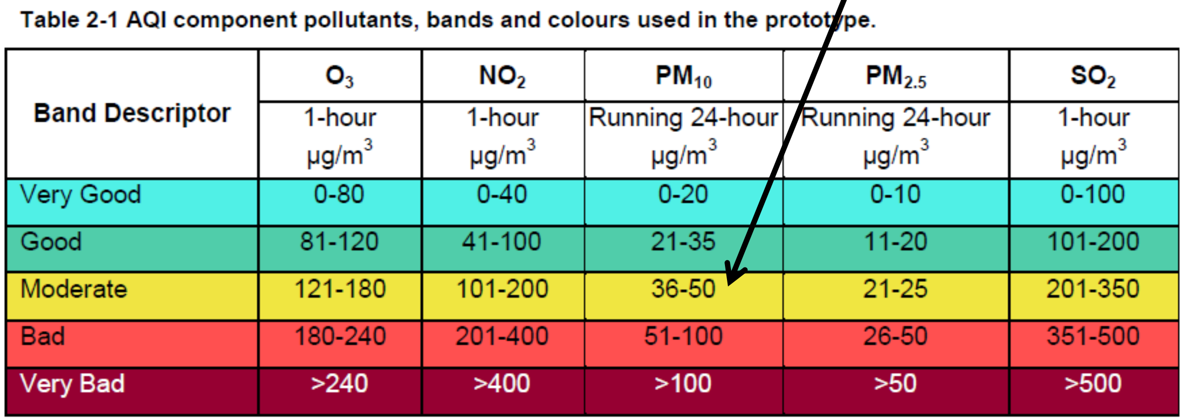


Table 2-1 AQI component pollutants, bands and colours used in the prototype.

Band Descriptor	O_3	NO_2	PM_{10}	$PM_{2.5}$	SO_2
	1-hour µg/m^3	1-hour µg/m^3	Running 24-hour µg/m^3	Running 24-hour µg/m^3	1-hour µg/m^3
Very Good	0-80	0-40	0-20	0-10	0-100
Good	81-120	41-100	21-35	11-20	101-200
Moderate	121-180	101-200	36-50	21-25	201-350
Bad	180-240	201-400	51-100	26-50	351-500
Very Bad	>240	>400	>100	>50	>500

The 'moderate' threshold for PM_{10} is 36 µg/m³. When this threshold is entered, DELTA outputs 'Moderate', 'Bad' and 'Very Bad' all together

Prototype EU Air Quality Index (2016)

(Ricardo report for DG ENV)

Threshold criteria

- What are we evaluating against i.e. what are our threshold criteria?
- These differ across Europe:
  - Threshold names
  - Threshold values
  - Index values
  - Pollutant averaging times

In the DELTA Tool:

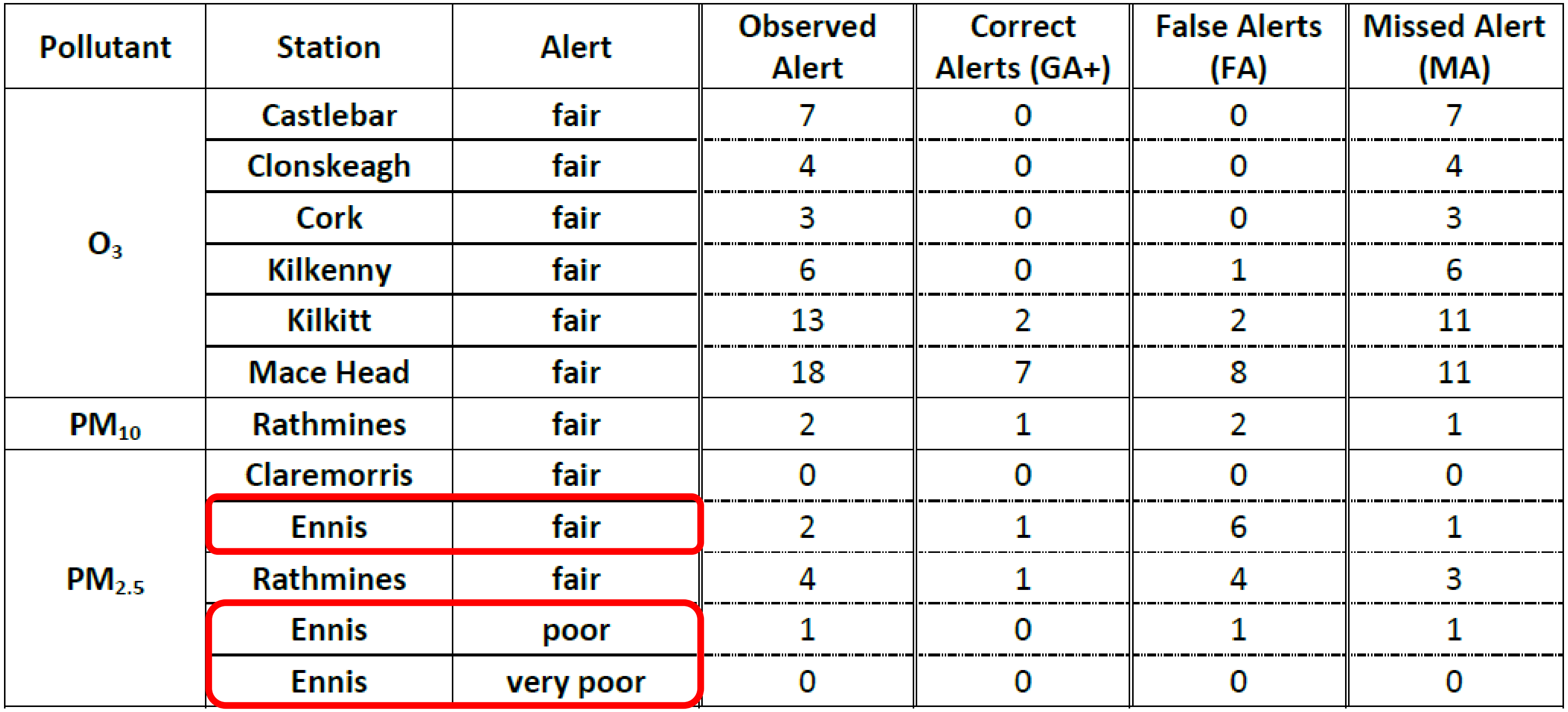
- Each pollutant is run separately
- Each threshold is entered separately
- A lower threshold will include the higher exceedance values e.g.

The 'moderate' threshold for PM_{10} is 36 µg/m³. When this threshold is entered, DELTA outputs 'Moderate', 'Bad' and 'Very Bad' all together

So until you know which pollutants have alerts, and what levels these are, you have to work through each pollutant and each threshold one by one…**very time consuming**

System evaluation

- What do we want to know to start with? **Summary statistics** (as output from the Model Evaluation Toolkit, no account of observation uncertainty):



Pollutant	Station	Alert	Observed Alert	Correct Alerts (GA+)	False Alerts (FA)	Missed Alert (MA)
O_3	Castlebar	fair	7	0	0	7
	Clonskeagh	fair	4	0	0	4
	Cork	fair	3	0	0	3
	Kilkenny	fair	6	0	1	6
	Kilkitt	fair	13	2	2	11
	Mace Head	fair	18	7	8	11
PM_{10}	Rathmines	fair	2	1	2	1
$PM_{2.5}$	Claremorris	fair	0	0	0	0
	Ennis	fair	2	1	6	1
	Rathmines	fair	4	1	4	3
	Ennis	poor	1	0	1	1
	Ennis	very poor	0	0	0	0

- Air quality generally good in Ireland, so few examples of cases where there are exceedances of the higher thresholds
- But in other areas e.g. London, there are many exceedances of these thresholds
  - Often more than one forecast per day (e.g. am, pm)

System evaluation

- What do we want to know to start with? **Summary statistics** (as output from the DELTA Tool in the dump file):

MO – mean observed

MM – mean modelled

SO – standard deviation observed

SM – standard deviation modelled

ExcO – observed exceedences

ExcM – modelled exceedences

GA+ – correct alerts

GA- – correct non-alerts

FA – false alerts

MA – missed alerts

CA – observed alerts

New for DELTA v5.5!

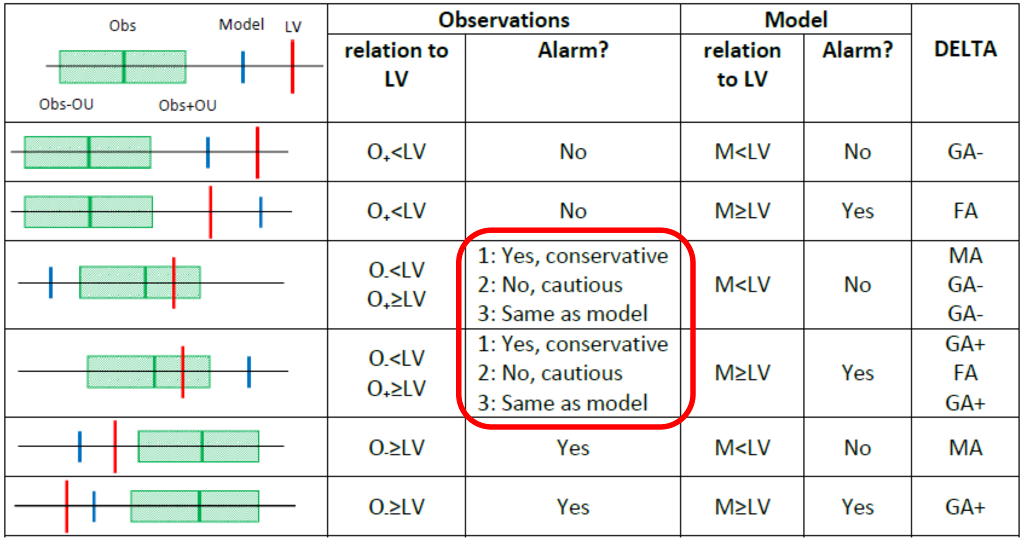
- Step in the right direction
- <u>But</u> you still have to process pollutants & thresholds separately – ideally at least all thresholds would be processed together

Note:

- ExcO & **CA** are the same for OU = 0
- When OU ≠ 0, ExcO stays as the OU = 0 value, but **CA** changes
- This may be fine, but the documentation does not say that ExcO doesn’t take into account OU

Flexibility options

- Which brings us on to the flexibility options:
  - ‘**Conservative**’ ~ assume there is an alert if there is a possibility there was
  - ‘**Cautious**’ ~ assume there isn’t an alert if there is a possibility there wasn’t
  - ‘**Same as model**’ ~ if there is uncertainty associated with whether or not there was an alert, then just opt for what the model indicates – may exaggerate the skill of the model



Obs Model LV Obs-OU Obs+OU	Observations		Model		DELTA
	relation to LV	Alarm?	relation to LV	Alarm?	
	$O_+<LV$	No	M<LV	No	GA-
	$O_+<LV$	No	M≥LV	Yes	FA
	$O_-<LV$ $O_+≥LV$	1: Yes, conservative 2: No, cautious 3: Same as model	M<LV	No	MA GA- GA-
	$O_-<LV$ $O_+≥LV$	1: Yes, conservative 2: No, cautious 3: Same as model	M≥LV	Yes	GA+ FA GA+
	$O_-≥LV$	Yes	M<LV	No	MA
	$O_-≥LV$	Yes	M≥LV	Yes	GA+

Table 1: Possible cases with respect with model, observation and associated uncertainty. Please note that some “<” or ”>” signs from the Note table have been changed to “≤” or ”≥” to make sure all situations are included (please check). The DELTA column indicates how DELTA considers the specific cases here described.

Note:

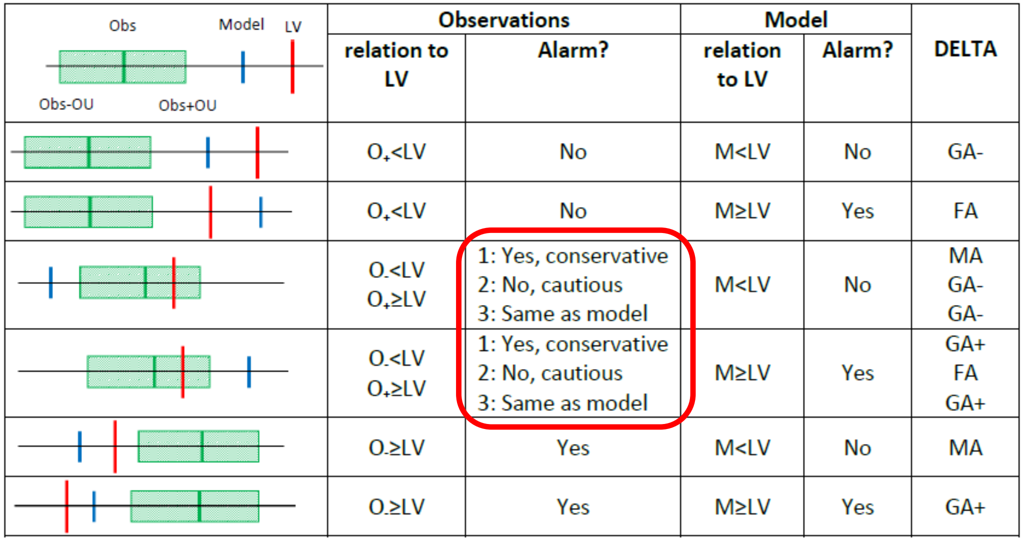
- ExcO & CA are the same for OU = 0
- When OU ≠ 0, ExcO stays as the OU = 0 value, but CA changes
- This may be fine, but the documentation does not say that ExcO doesn’t take into account OU

CERC

Flexibility options

- CERC suggested:
  - ‘**Certain**’ ~ restrict the assessment to those data points where it is certain that an alert was or was not exceeded
  - We are **not** suggesting that ‘Certain’ is the same as setting OU = 0 (as stated in .doc)
  - ‘Certain’ should be a valid option for all values of OU, it should just **exclude** the cases where

    LV ∊ [Obs-OU,Obs+OU]

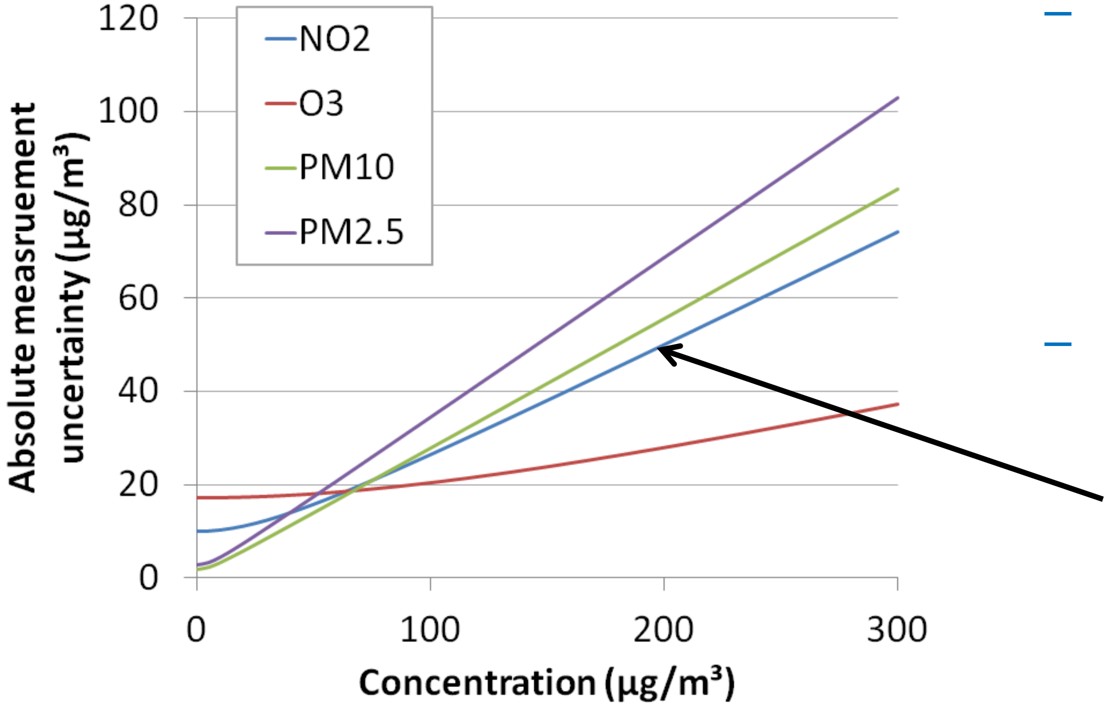


Obs Model LV Obs-OU Obs+OU	Observations		Model		DELTA
	relation to LV	**Alarm?**	**relation to LV**	**Alarm?**	
	$O_+<LV$	No	M<LV	No	GA-
	$O_+<LV$	No	M≥LV	Yes	FA
	$O_-<LV$ $O_+\geq LV$	1: Yes, conservative 2: No, cautious 3: Same as model	M<LV	No	MA GA- GA-
	$O_-<LV$ $O_+\geq LV$	1: Yes, conservative 2: No, cautious 3: Same as model	M≥LV	Yes	GA+ FA GA+
	$O_-\geq LV$	Yes	M<LV	No	MA
	$O_-\geq LV$	Yes	M≥LV	Yes	GA+

Table 1: Possible cases with respect with model, observation and associated uncertainty. Please note that some “<” or ”>” signs from the Note table have been changed to “≤” or ”≥” to make sure all situations are included (please check). The DELTA column indicates how DELTA considers the specific cases here described.

Flexibility options

- CERC suggested:
  - ‘**Certain**’ ~ restrict the assessment to those data points where it is certain that an alert was or was not exceeded
  - We are **not** suggesting that ‘Certain’ is the same as setting OU = 0 (as stated in .doc)

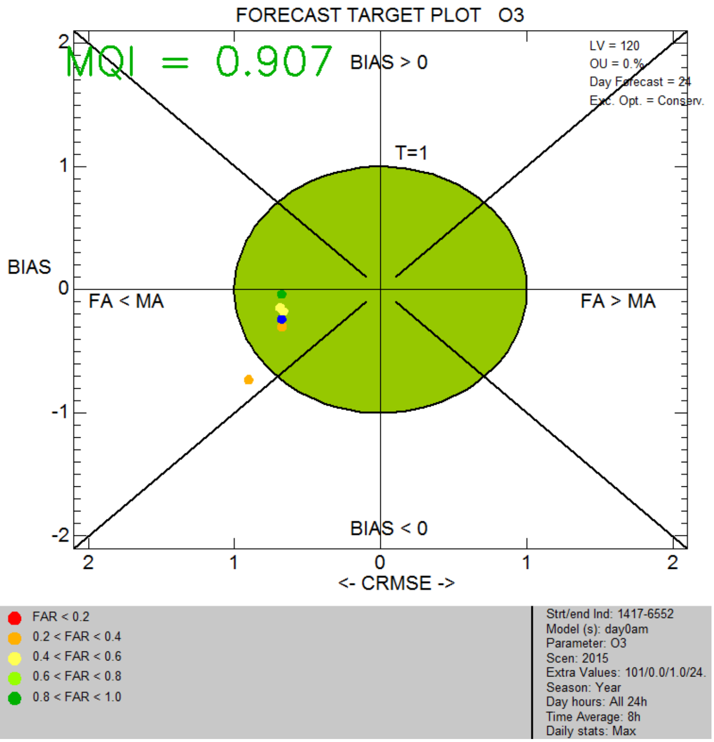


- ‘Certain’ should be a valid option for all values of OU, it should just **exclude** the cases where

  LV ∊ [Obs-OU,Obs+OU]
- This may be problematic - measurement uncertainties are large when concentrations are high i.e. at the threshold values

Items ‘to be discussed at meeting’

- ‘4. *It would be helpful to give guidance on whether or not fixed values or variable values of OU should be used*.’
  - Default is Assessment uncertainty, other OU to be introduced as expert users ✓
- *‘7 a. When assessing a forecast, isn’t the most important point how good the system is at accurately producing an alert? A possible issue with the target diagram is that it appears to focus on the target rather than the system’s ability to predict alerts.’*
  - Think about a possible summary report including additional indicators e.g. GA+, GA-, FA, MA **– to discuss**



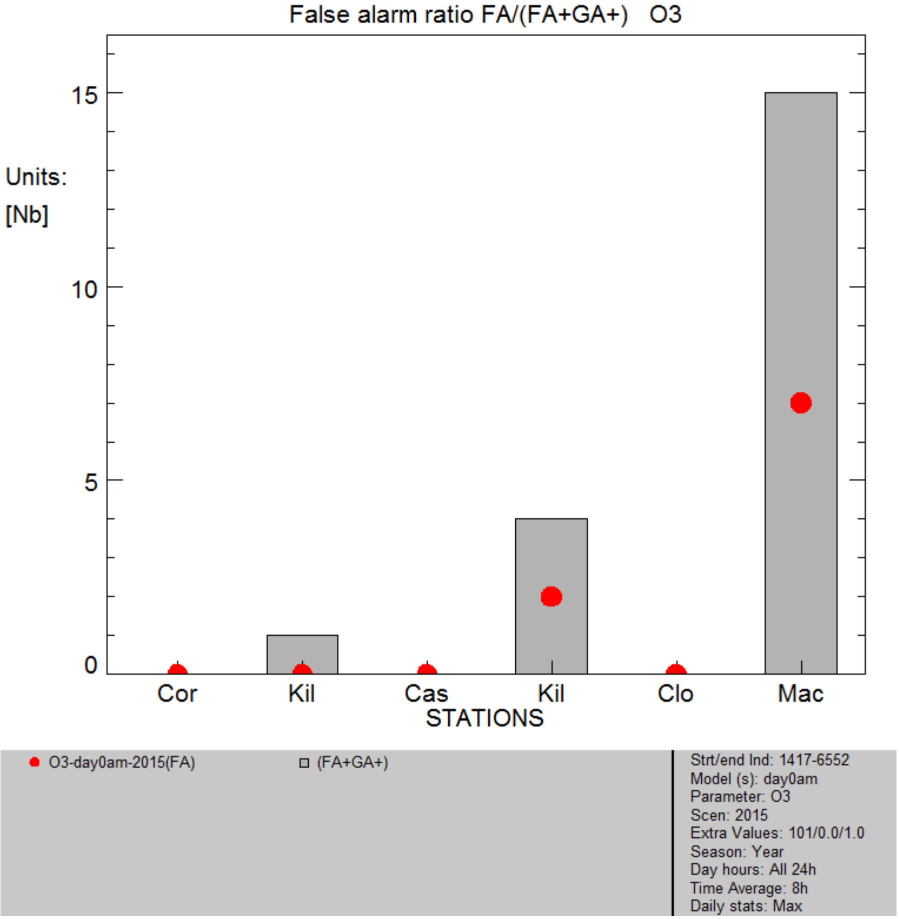
LV = 120
OU = 0.%
Day Forecast = 24
Exc. Opt. = Conserv.

Strt/end Ind: 1417-6552
Model (s): day0am
Parameter: O3
Scen: 2015
Extra Values: 101/0.0/1.0/24.
Season: Year
Day hours: All 24h
Time Average: 8h
Daily stats: Max

CERC

Items ‘to be discussed at meeting’

- *‘15 a. False Alarm Ratio plot*
  - *Red spot is the number of correct alerts (GA+), grey bar is the number of correct alerts plus false alarms (GA+ + FA), i.e. grey bar shows how many alerts were issued and the red spot how many were correct.*
  - *Title is misleading’*
  - Title says:
    - “False alarm ratio plot FA/(FA+GA+) O3”
    - **But the plot axis is not a ratio**
    - Should say something like “Comparison of correct model alerts with total model alerts”
  - Similar issue for Probability of Detection plot
  - Philippe says he updated?



Strt/end Ind: 1417-6552
Model (s): day0am
Parameter: O3
Scen: 2015
Extra Values: 101/0.0/1.0
Season: Year
Day hours: All 24h
Time Average: 8h
Daily stats: Max

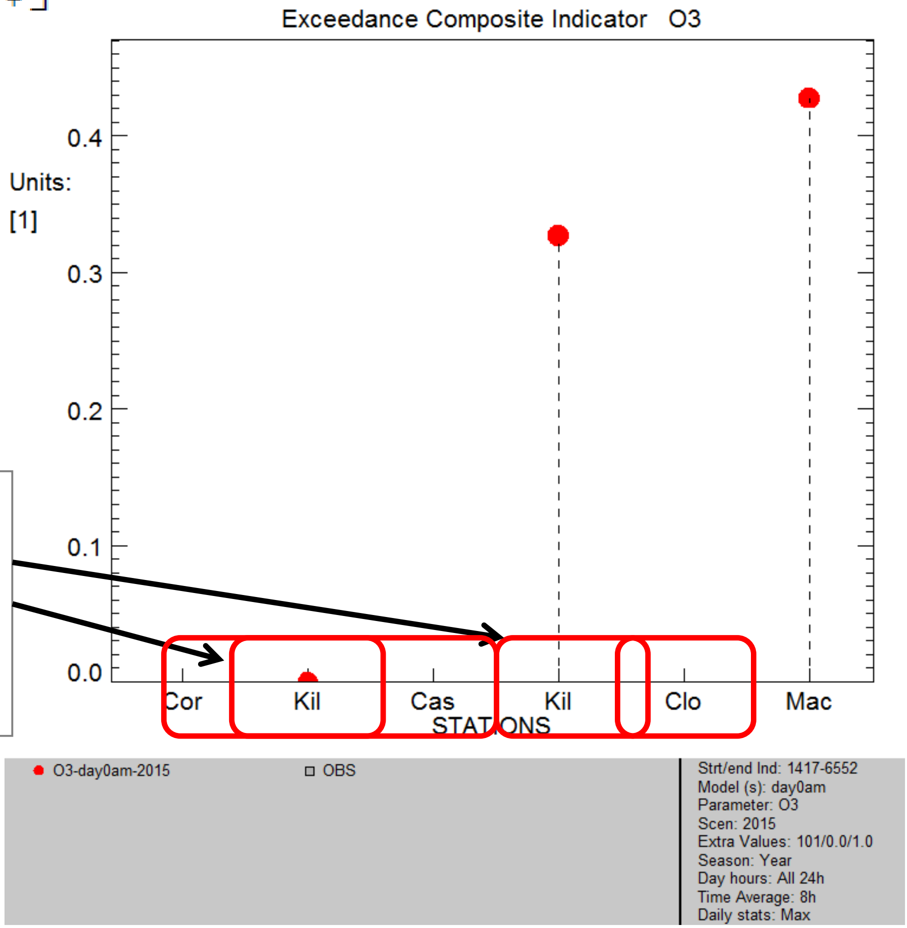
Items ‘to be discussed at meeting’

- *’15 d. Exceedence Indicator*
  - *The red spot is the ratio:*

$$0.5\left[\frac{GA_+}{MA+GA_+}+\frac{GA_+}{FA+GA_+}\right]$$

  - *This needs more thought because of the NaN when, e.g. FA+GA+=0*
  - *Also, need to indicate in legend why some points are not shown’* i.e. NAN issue

Also, only using the first three letters of the station name means that ‘Kilkenny’ and ‘Kilkitt’ are indistinguishable



Strt/end Ind: 1417-6552
Model (s): day0am
Parameter: O3
Scen: 2015
Extra Values: 101/0.0/1.0
Season: Year
Day hours: All 24h
Time Average: 8h
Daily stats: Max

CERC

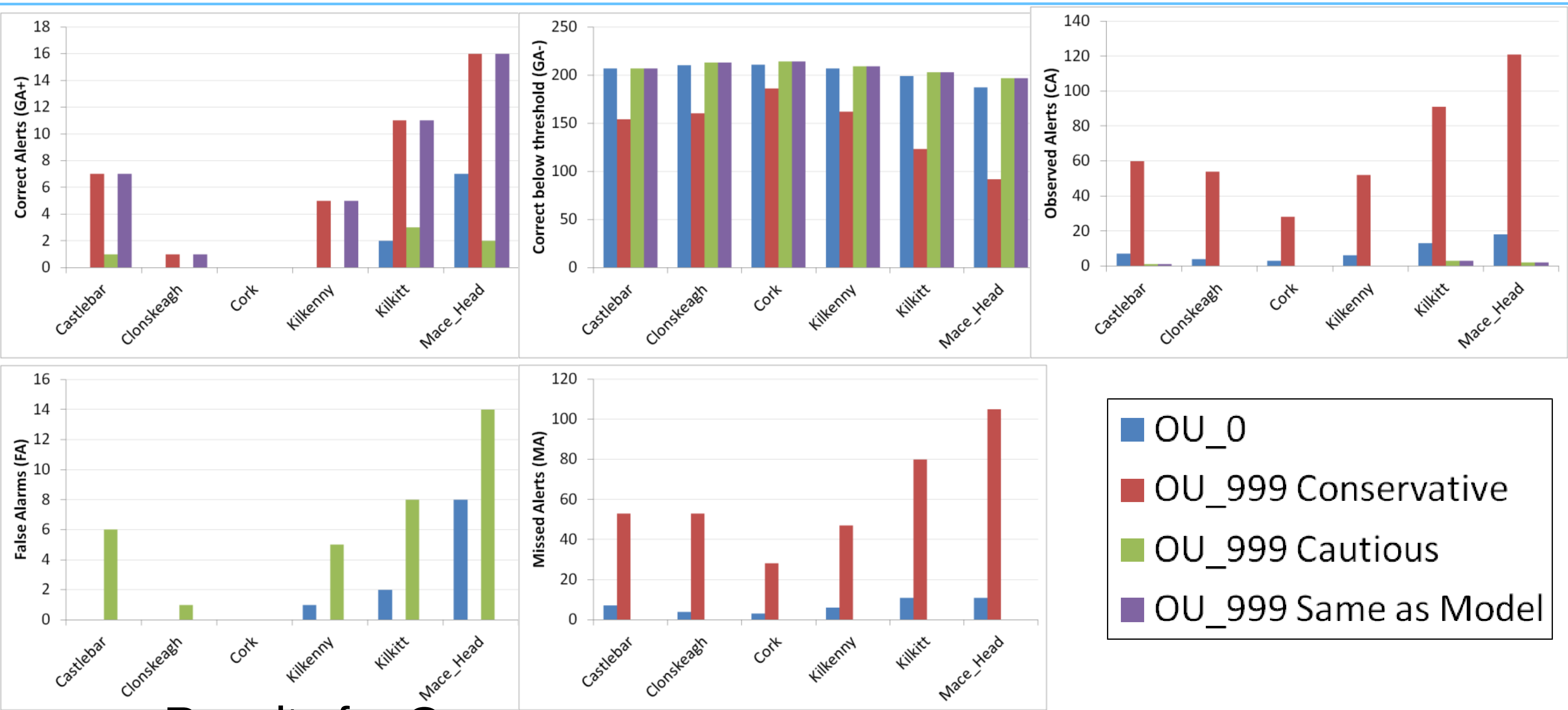
Summary

- There have been some improvements to the forecasting mode of the DELTA tool
- Using the tool for a ‘real’ project highlighted some issues with usability, particularly:
  - relating to the number of times you have to run the tool (i.e. no. of forecasts x no. of pollutants x no. of thresholds and/or indices)
  - its flexibility with respect to the different European threshold criteria (e.g. pollutant averaging times)
- The best way to account of observation uncertainty for these assessments is still not clear

- If time during the meeting, it would be good to resolve the ‘Remaining issues’ (Section 5 of document) as some of these are out of date & we should possibly add new ones?

Additional slides

Flexibilty options & GA+, GA-, MA, FA, CA



- Results for O_3
  - ‘Conservative’ means that there are many alerts, and many missed alerts
  - ‘Cautious’ means that there aren’t many alerts so quite a few false alarms
  - For this case ‘same as model’ gives FA = MA = 0 i.e. perfect!