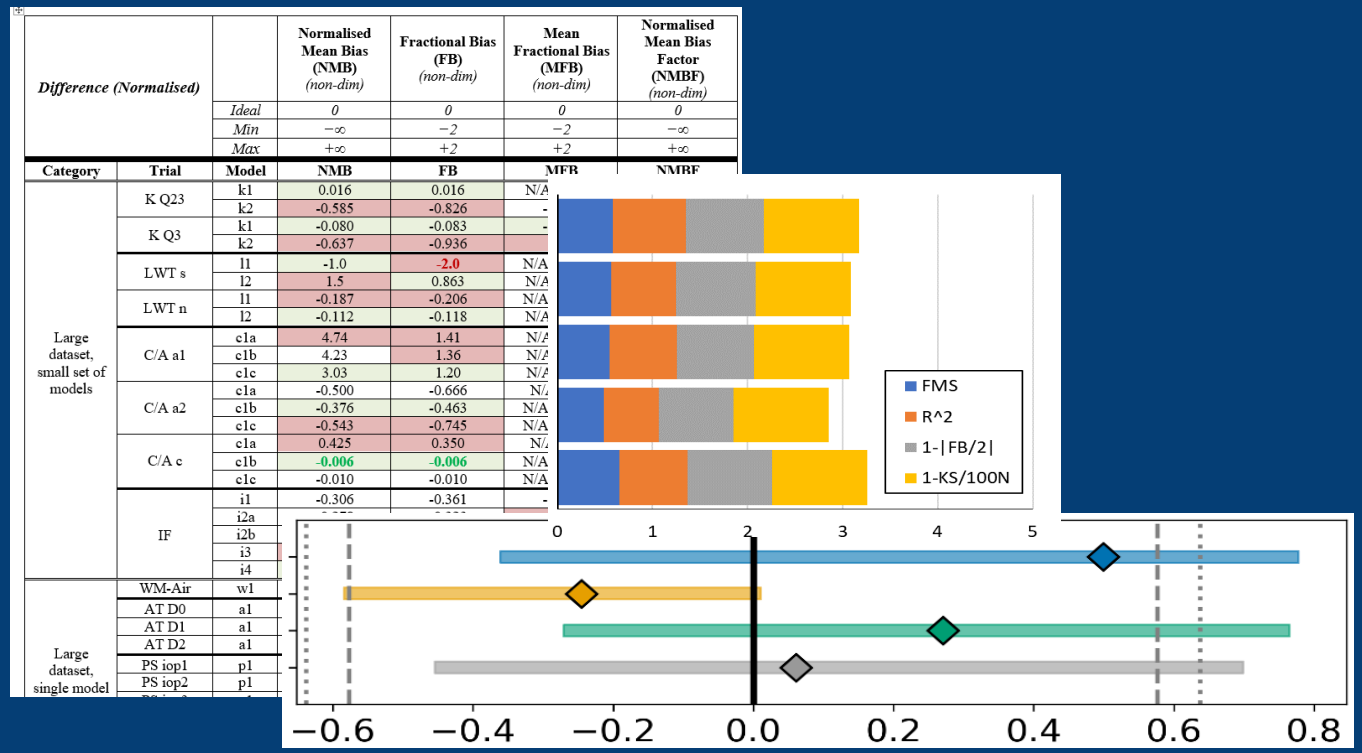
A review of methods used to assess the performance of atmospheric dispersion models: overview

Jenny Stocker



Difference (Normalised)			Normalised Mean Bias (NMB) *(non-dim)*	Fractional Bias (FB) *(non-dim)*	Mean Fractional Bias (MFB) *(non-dim)*	Normalised Mean Bias Factor (NMBF) *(non-dim)*
		Ideal	*0*	*0*	*0*	*0*
		Min	*−∞*	*−2*	*−2*	*−∞*
		Max	*+∞*	*+2*	*+2*	*+∞*
Category	**Trial**	**Model**	**NMB**	**FB**	**MFB**	**NMBF**
Large dataset, small set of models	K Q23	k1	0.016	0.016	N/A	
		k2	-0.585	-0.826	-	
	K Q3	k1	-0.080	-0.083	-	
		k2	-0.637	-0.936		
	LWT s	l1	-1.0	-2.0	N/A	
		l2	1.5	0.863	N/A	
	LWT n	l1	-0.187	-0.206	N/A	
		l2	-0.112	-0.118	N/A	
	C/A a1	c1a	4.74	1.41	N/A	
		c1b	4.23	1.36	N/A	
		c1c	3.03	1.20	N/A	
	C/A a2	c1a	-0.500	-0.666	N/A	
		c1b	-0.376	-0.463	N/A	
		c1c	-0.543	-0.745	N/A	
	C/A c	c1a	0.425	0.350	N/A	
		c1b	-0.006	-0.006	N/A	
		c1c	-0.010	-0.010	N/A	
	IF	i1	-0.306	-0.361	-	
		i2a				
		i2b				
		i3				
		i4				
Large dataset, single model	WM-Air	w1				
	AT D0	a1				
	AT D1	a1				
	AT D2	a1				
	PS iop1	p1				
	PS iop2	p1				

Model Evaluation Workshop: Thresholds & Toolkit

UKHSA Training Centre

11th June 2026

ADMLC

CERC

Presentation outline

- Introduction
  - Study motivation & aims
  - Project (report) overview
  - Literature review
  - Model Evaluation Toolkit
- Technical summary (case studies & metrics)
- Results (statistical & graphical)
- Analyses (metrics thresholds, error analyses)
- Metric suitability (different datasets / applications)
- Recommendations



Up to 5 minutes questions



Up to 5 minutes questions



Remaining questions

Introduction: study motivation & aims

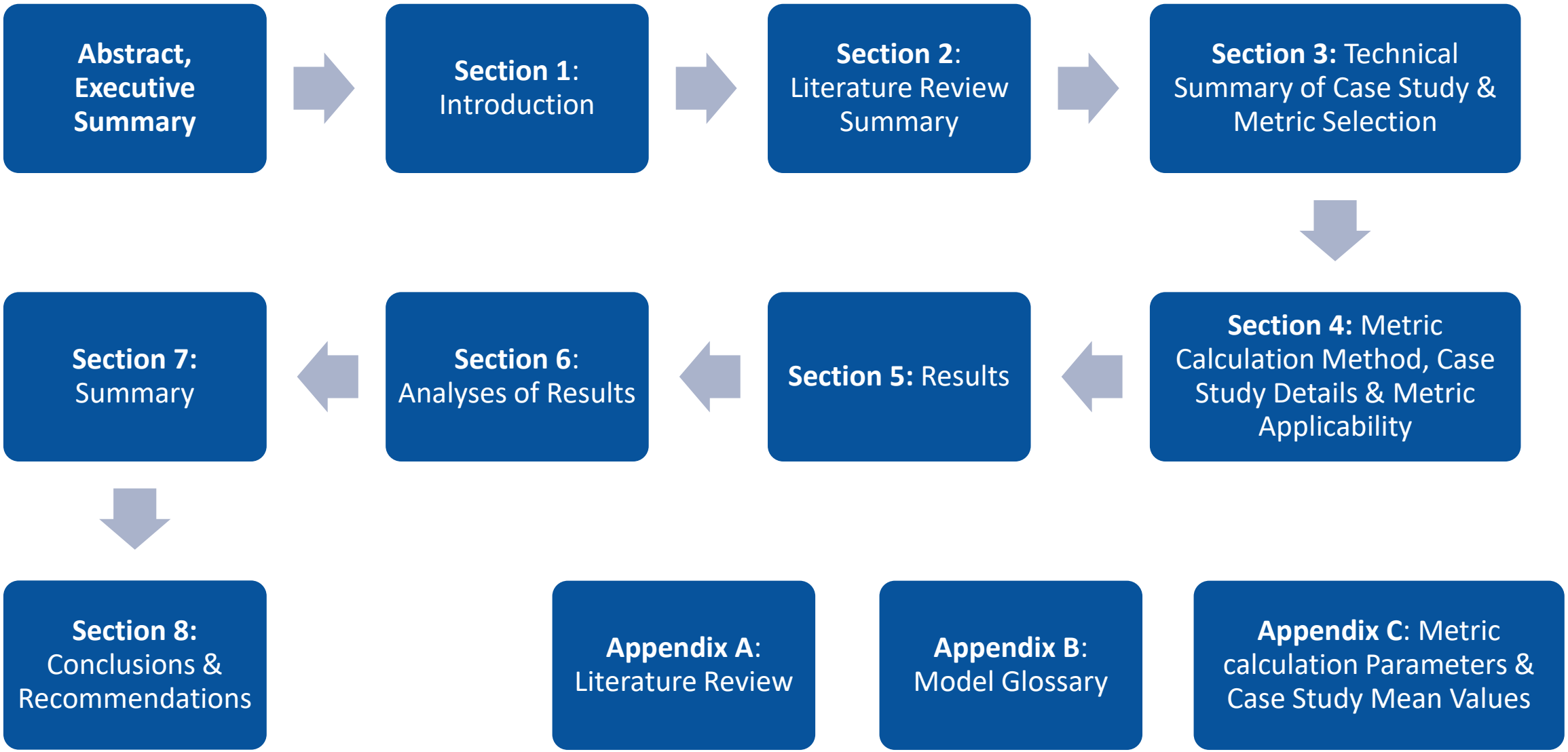
Study motivation

- Numerical atmospheric dispersion models require evaluation using measurement datasets
- Many different evaluation methods
- The majority of methods include calculation of different metrics and graphics
- Modellers would benefit from guidance on:
  - Metrics available
  - Useful graphics

Study aims

- Collate metric and graphic lists (literature review, ADMLC opinions)
- Collate case study datasets (modelled & measured datasets) covering range of atmospheric dispersion model applications
- Modify CERC's Model Evaluation Toolkit to generate required metrics
- Metrics:
  - Classify according to type
  - Document ideal values & possible ranges
  - Analyse calculated values to generate thresholds
  - Describe use of metrics to analyse model error
  - Assess metrics' suitability for different datasets and applications
- Demonstrate usefulness of graphics
- Guidance for model users

Introduction: project (report) overview



Introduction: literature review summary

- Detailed literature review, with content reviewed by Steve Hanna
- Considers topics including:
  - Motivation for model evaluation
  - Atmospheric dispersion model applications
  - Model types
  - Observed data types suitable for model evaluation & associated challenges
  - Model evaluation tools
  - Metrics & graphics used in the literature
- Categorises and defines metrics
- Presents example graphics from the literature

Table 9.9 – Model performance metrics based on the areas in which models and observations indicate

Metric	Equation	Value for ideal model	Description
AFN	$A_o - (A_p \cap A_o)$ Or pairwise (using data summation as an estimate for area) $\frac{1}{2}\sum_i[\lvert C_{oi} - C_{pi}\rvert + (C_{oi} - C_{pi})]$	0	Area of false negatives For a particular threshold value, AFN is the area over which exceedance is observed but not predicted
AFP	$A_p - (A_p \cap A_o)$ Or pairwise (using data summation as an estimate for area) $\frac{1}{2}\sum_i[\lvert C_{oi} - C_{pi}\rvert + (C_{pi} - C_{oi})]$	0	Area of false positives For a particular threshold value, AFP is the area over which exceedance is predicted but not observed
FMS	$\frac{A_p \cap A_o}{A_p \cup A_o}$ where A_p is the predicted contour area based on a certain threshold A_o is the observed contour area based on the same threshold Note: Mosca et al. (1998) includes a factor of 100 in the definition, to express in % Or pairwise: $\frac{\frac{1}{2}\sum_i[(C_{oi} + C_{pi}) - \lvert C_{oi} - C_{pi}\rvert - C_t}{A_p \cap A_o + A_{FN} + A_{FP}}$	1	Figure of merit in space (also known as threat score) For a particular threshold value, FMS is the area over which exceedance is correctly predicted as a fraction of the area where exceedance is predicted and/or observed, at a fixed time For the pairwise exceedance comparison, a minimum concentration threshold C_t must be defined so that exceedances represent concentrations meaningfully above measurement noise and to ensure numerical stability in the exceedance-based

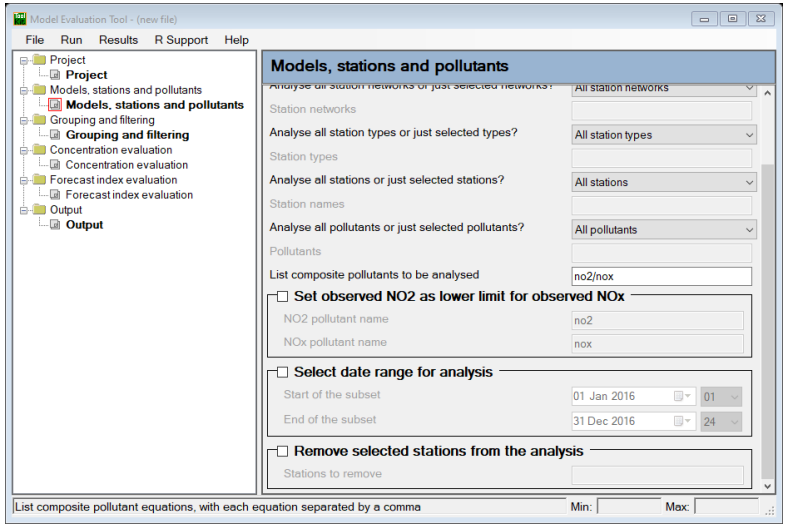
Introduction: metric summary

- Metrics are broadly categorised by ‘Metric type’ and what ‘applications’ different metric types are suitable for

Metric type	Metric type description	Metric type application(s)	Example metric
Difference (Absolute)	Absolute differences	General purpose	Mean Bias (MB)
Difference (Normalised)	Relative normalised differences	General purpose	Normalised Mean Bias (NMB)
Error (Absolute)	Absolute errors	General purpose	Mean Error (ME)
Error (Normalised)	Relative normalised errors	General purpose	Normalised Absolute Error (NAE)
Fractional	Proportion within a threshold	General purpose	Factor of 2 (FAC2)
Linearity	Linearity of the relationship	General purpose	Correlation coefficient (R)
Area Exceedances	Exceedance of areas of a threshold	Spatial performance	Figure of Merit in Space (FMS)
Peak Concentrations	Peak values	Peak concentrations	Accuracy of Unpaired peak (AU)
FAIRMODE	Developed by FAIRMODE initiative	EU air quality directive	Model Quality Index (MQI)
Forecasting	Forecasting metrics	Air quality forecasting	Probability of Detection (POD)
Summary/combination	Comprised of multiple metrics	Model intercomparison	Index of Agreement (IoA)

Introduction: Model Evaluation Toolkit

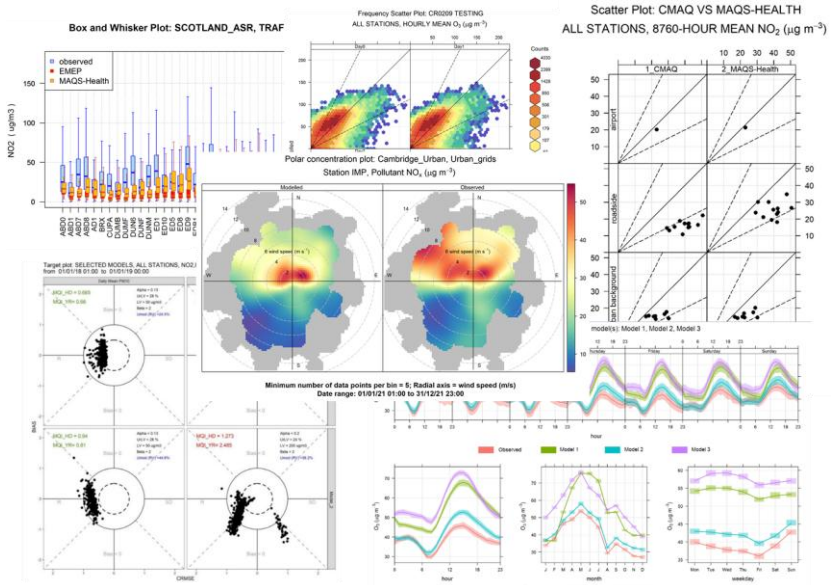
Model Evaluation Toolkit	Input	Output
Data Input Tool	• Flexible input for modelled data • For this project all data was re-formatted into “hourly” CSV data	R-workspace of observed and modelled data
Model Evaluation Tool	R-workspace User defined grouping and filtering	• Graphs • CSV statistics



	A	B	C	D	E	F
1	station	pollutant	model	type	output.avg.time	num.valid
2	abd	no2	MAQS	urban background	annual mean	1
3	abd0	no2	MAQS	roadside	annual mean	1
4	abd1	no2	MAQS	roadside	annual mean	1
5	abd3	no2	MAQS	roadside	annual mean	1
6	abd8	no2	MAQS	roadside	annual mean	1

	A	B	C	D	E
1	type	pollutant	model	output.avg.time	num.valid
2	industrial	pm2.5	MAQS	annual mean	14
3	roadside	pm2.5	MAQS	annual mean	64
4	rural background	pm2.5	MAQS	annual mean	5
5	urban background	pm2.5	MAQS	annual mean	42

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	station	pollutant	model	output.avg.time	num.valid	obs.mean	mod.mean	SDO	SDM	MB	NMSE	R	R2	Fac2	Fb
2	all	no2	Urban	hourly mean	8172	16.28	14.57	11.61	10.26	-1.71	0.29	0.73	0.53	0.82	-0.11
3	all	nox	Urban	hourly mean	8172	28.16	26.12	25.60	22.31	-2.04	0.40	0.76	0.57	0.80	-0.08



Technical summary: case studies

Case Study	Use	Categorisation	Models	Time periods	Obs. sites	Range (km)	Data
Kincaid (K)	Regulatory	Large dataset, small set of models	2	~60 hours	~5 arcs (sensors)	50	Arc max
Lee Wind Tunnel (LWT)	Regulatory	Large dataset, small set of models	2	156 / 63 cases	6 arcs (sensors)	<1	Arc max
CAPTEX / ANATEX (C/A)	Emergency response	Large dataset, small set of models	3*	15 / 84 & 11 measurements	23 / ~75 (sampling stations)	800 / 3000	3-hour avg / 24-hour avg
Idaho Falls (IF)	Regulatory (road)	Large dataset, small set of models	5*	4 trials	56 (bag samplers)	0.2	15-min avg
WM-Air	Compliance (urban)	Large dataset, single model	1	8784 hours	32 (monitors)	50	Cont. hrly
airTEXT (AT)	Operational forecast (urban)	Large dataset, single model	1	8760 hours	68 (monitors)	60	Cont. hrly
Project Sagebrush 1 (PS)	Regulatory (industrial)	Large dataset, single model	1	5 trials	5 arcs (~1350 bag samplers)	3.2	10-min avg
Desert Tortoise & FLADIS (DT&F)	Accidental release	Small dataset, large set of models	19	3 trials	2 / 3 arcs (sensors)	0.8 / 0.24	Avg. arc max
Jack Rabbit II (JRII)	Accidental release	Small dataset, large set of models	16	3 trials	6 arcs (sensors)	11	Avg. arc max

Note: table refers to data included for the analyses in this report, additional data may be available.

Technical summary: metric selection

- Some metrics were excluded because:
  - rarely used in literature
  - depend on a metric that has been excluded
  - similar to another metric
  - give invalid values if observed and/or modelled values are zero
- Example of a metric that behaves badly for low values of observed concentration:
  Mean Normalised Error (MNE)

C_p Modelled concentrations

C_o Observed concentrations

$$\text{MNE} = \overline{\left|\frac{C_p - C_o}{C_o}\right|}$$

Average is calculated after normalisation of each term

If $C_o \ll 1$ then the difference between C_p and C_o for that data pair is 'amplified'

- This study considered broadly applicable metrics, so metric behaving in this way have been excluded
- For datasets that do not have $C_o \ll 1$, such metrics can be calculated and would be 'well behaved'

Technical summary: metric selection

- Metrics included in final analysis:

Metric Type	Metrics
Difference (Absolute)	MB
Difference (Normalised)	NMB, FB, NMBF
Error (Absolute)	MSE, MSEs, MSEu, MSE.cor, mMSE, ME, MAE, RMSE, CRMSE
Error (Normalised)	NMSE, NMSEs, NMSEu, NMSE.cor, mNMSE, NAE, HH, NSD, NRMSE, NCRMSE, NMSD, Fs
Fractional	FAC2, FACn, AE band
Linearity	r, rrank, θ
Area Exceedances	A_{FN}, A_{FP}, FMS, 1D MOE, 2D MOE
Peak Concentrations	AU, AP, τ, ε, FMT, RHC
FAIRMODE	MQI, MQI annual, MPI bias, MPI r, MPI sd, Spatial MPI r, Spatial MPI sd
Forecasting	POD, SR, FBS, ACC, FAR, POFD, TS, GSS, ORSS, HSS, PSS, MPI FP, RANK2
Summary / combination	IoA, CoE, KS, RANK, ρ2, λ

Selection of metric types for case studies

- *Difference, Error* and *Peak* have no restrictions: **all case studies**
- *Fractional, Linearity, & Summary / Combination* are not suitable for very small datasets: **all except DT&F**
- *Area Exceedances*, applicable only to sufficiently large and well-spatially distributed datasets: **only LWT, C/A** & **IF**
- *FAIRMODE, Forecasting* require good quantity of good quality data: **only WM-Air** and **AT**

	Case study								
Metric type	**K**	**LWT**	**C/A**	**IF**	**WM - Air**	**AT**	**PS**	**DT & F**	**JRII**
Difference (Absolute)									
Difference (Normalised)									
Error (Absolute)									
Error (Normalised)									
Fractional									
Linearity									
Area Exceedances									
Peak Concentrations									
FAIRMODE									
Forecasting									
Summary / Combination									

Results: example table of statistics

- Results tables for all metrics grouped by metric type, e.g. ***Difference (Normalised)***
- Case studies are grouped by **dataset categorisation**
- Colour coding *explained later*

Difference (Normalised)			Normalised Mean Bias (NMB) *(non-dim)*	Fractional Bias (FB) *(non-dim)*	Mean Fractional Bias (MFB) *(non-dim)*	Normalised Mean Bias Factor (NMBF) *(non-dim)*
		Ideal	*0*	*0*	*0*	*0*
		Min	*−∞*	*−2*	*−2*	*−∞*
		Max	*+∞*	*+2*	*+2*	*+∞*
Category	**Trial**	**Model**	**NMB**	**FB**	**MFB**	**NMBF**
Large dataset, small set of models	K Q23	k1	0.016	0.016	N/A (0.043)	0.016
		k2	-0.585	-0.826	-0.809	-1.4
	K Q3	k1	-0.080	-0.083	-0.087	-0.086
		k2	-0.637	-0.936	-1	-1.8
	LWT s	l1	-1.0	-2.0	N/A (-0.242)	-8.93×10^{5}
		l2	1.5	0.863	N/A (0.672)	1.5
	LWT n	l1	-0.187	-0.206	N/A (-0.201)	-0.230
		l2	-0.112	-0.118	N/A (0.017)	-0.126
	C/A a1	c1a	4.74	1.41	N/A (-1.08)	4.74
		c1b	4.23	1.36	N/A (-0.58)	4.23
		c1c	3.03	1.20	N/A (-0.67)	3.03
	C/A a2	c1a	-0.500	-0.666	N/A (-1.2)	-0.998
		c1b	-0.376	-0.463	N/A (-0.623)	-0.602
		c1c	-0.543	-0.745	N/A (-0.830)	-1.2
	C/A c	c1a	0.425	0.350	N/A (-1.3)	0.425
		c1b	-0.006	-0.006	N/A (-0.821)	-0.006
		c1c	-0.010	-0.010	N/A (-0.323)	-0.010
	IF	i1	-0.306	-0.361	-0.178	-0.441
		i2a	-0.278	-0.323	-0.458	-0.386
		i2b	-0.319	-0.380	-0.33	-0.469
		i3	-0.346	-0.418	-0.365	-0.529
		i4	-0.207	-0.230	-0.127	-0.260
Large dataset, single model	WM-Air	w1	-0.038	-0.039	-0.023	-0.039
	AT D0	a1	0.526	0.416	0.434	0.526
	AT D1	a1	0.511	0.407	0.427	0.511
	AT D2	a1	0.488	0.393	0.417	0.488
	PS iop1	p1	7.1	1.6	0.253	7.1
	PS iop2	p1	0.862	0.602	0.989	0.862

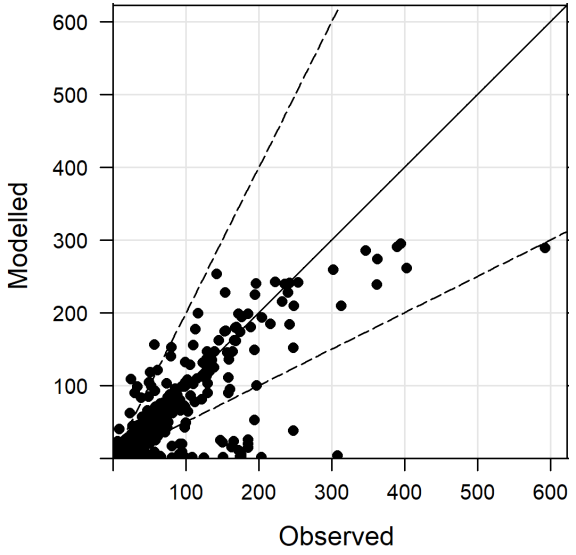
- Specify: **units** or is **non-dimensional**, **ideal value**, **min** & **max** possible values
- N/A:
  - N/A values indicate a **zero-value error**
  - (value) obtained by **removing zero value pairs** from the dataset
- *Italics* highlight a non-obvious power of 10 change e.g.

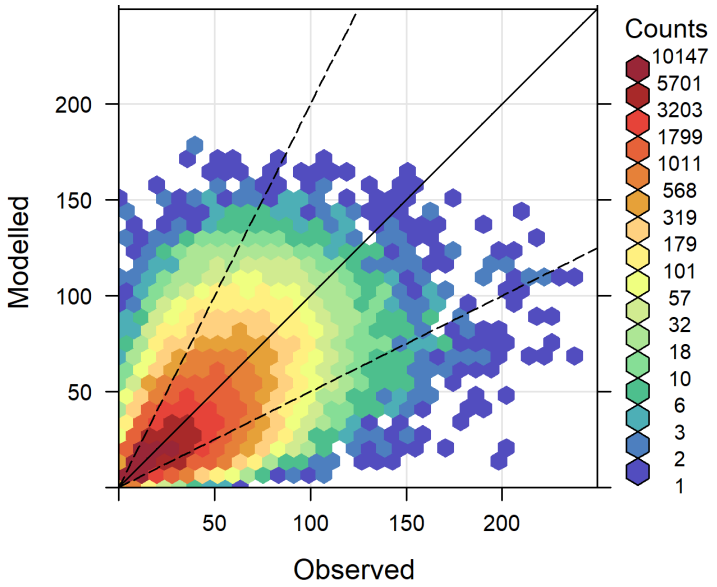
	0.03×10^{16}
a	0.05×10^{16}
b	1.23×10^{16}
a	2.84×10^{15}
b	0.01×10^{16}

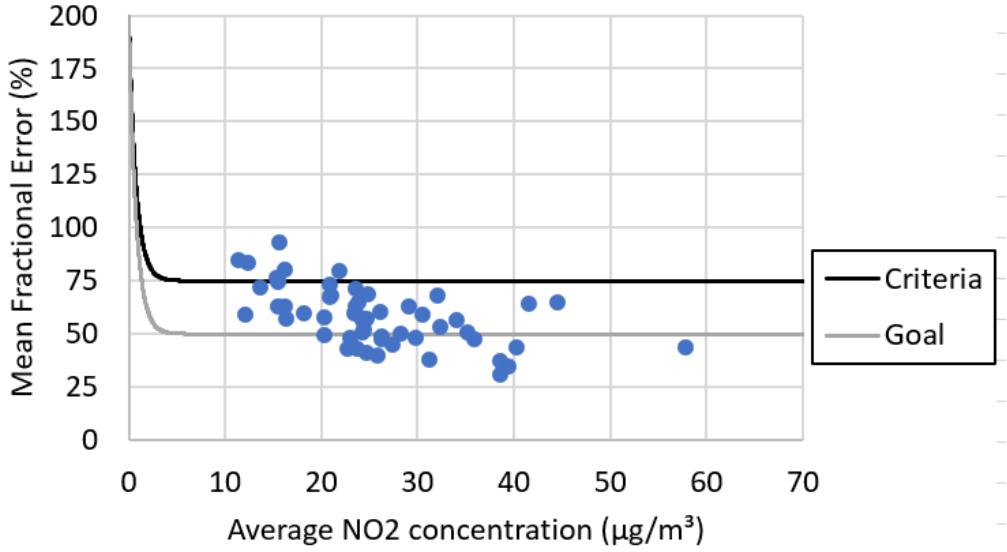
Results: graphical analysis of single model, single metric

- Useful for quickly and intuitively assessing and intercomparing performance

Use case	Graphical plot(s)	Examples	
Single model, single metric	• Frequency scatter plot • Bugle plot e.g. MFE, MFB • Conventional scatter plot • Quantile-quantile plot • Box-and-whisker plot	**Scatter plots** • Standard scatter plot • Frequency scatter plot • Very useful for ball-park checks & checking against limits e.g. factor of 2 • Should always be square!	**Bugle plot**: • named for the shape of the inverse relationship between the fractional metric and observations. Can define thresholds. • MFB goal ±30% and criteria ±60% shown for *air*TEXT



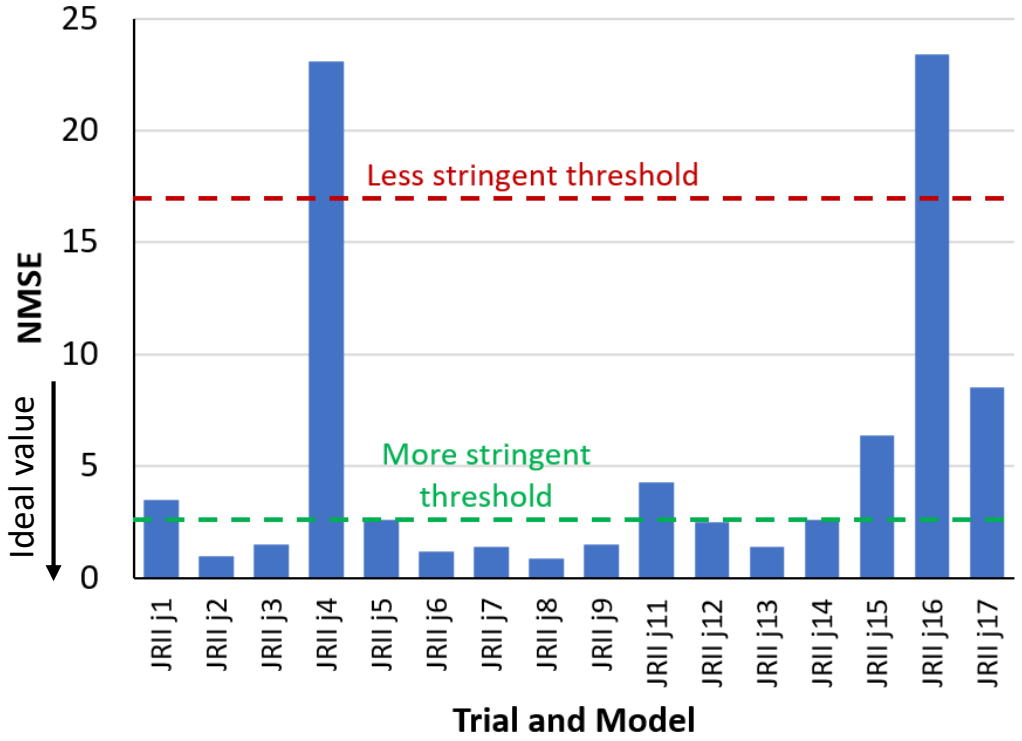


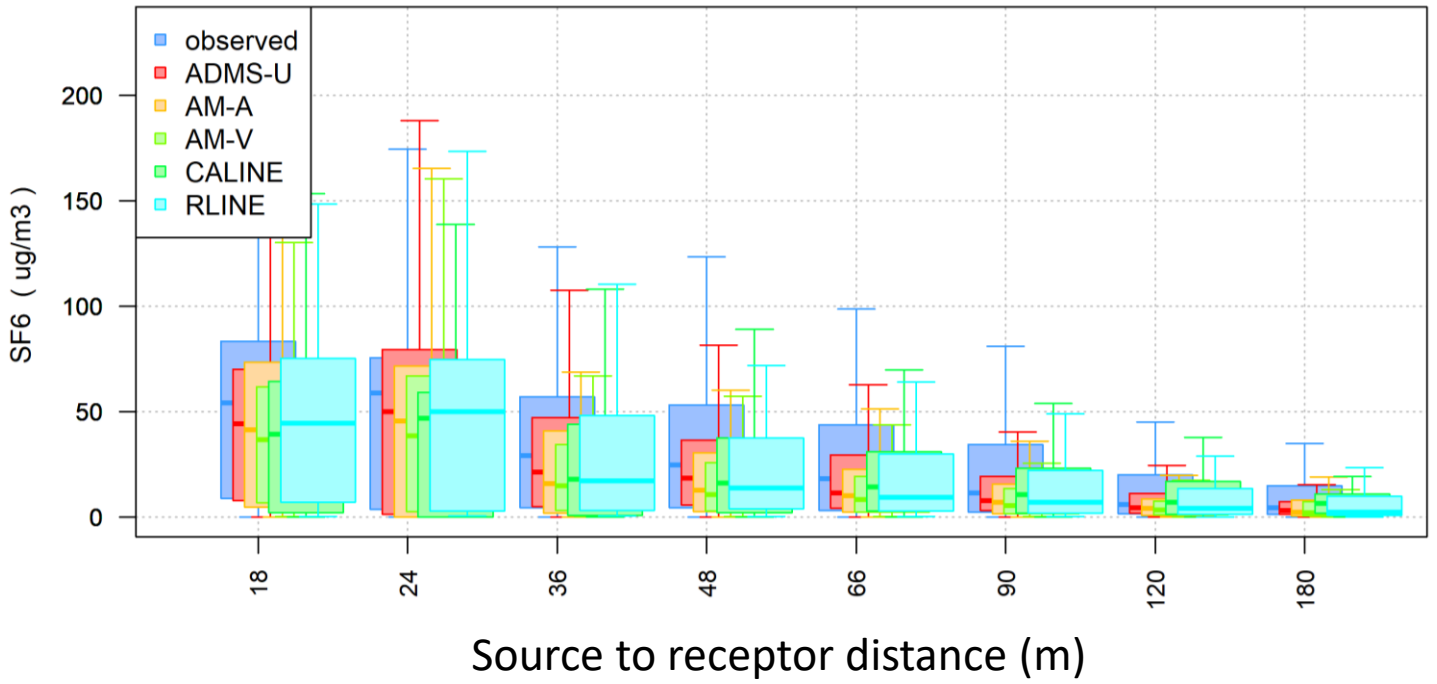


Results: graphical analysis of multi-model, single

- Useful for quickly and intuitively assessing and intercomparing performance

Use case	Graphical plot(s)	Examples	
Multi-model, single metric	• Bar charts of normalised metrics e.g. NMSE • Box-and-whisker plot	**Bar charts of normalised metrics:** • Simple • Easy comparison against threshold values • Clear outliers	**Box-and-whisker plot:** • Summarises distributions and compares interquartile ranges • Split by distance for each model for Idaho Falls

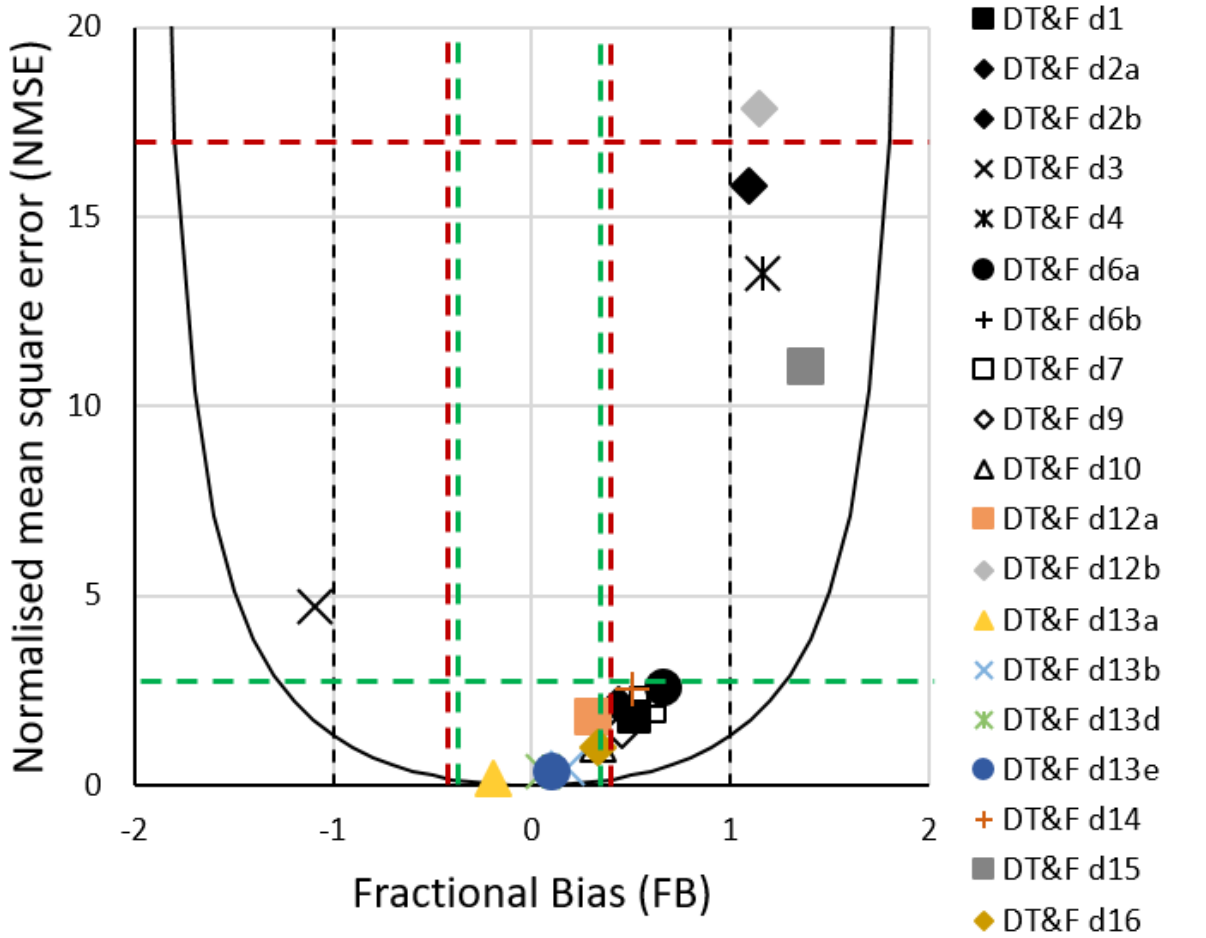




Results: graphical analysis multi-metric

- Useful for quickly and intuitively assessing and intercomparing performance

Use case	Graphical plot(s)	Example
Multi-metric, single- or multi-model	• Parabola plot of normalised metrics e.g. FB vs NMSE and MG vs VG • Target plot e.g. FAIRMODE target plot for MQO • Bar charts of decomposed metrics e.g. MSE, RANK • Taylor diagram	Parabola plot: • Compare model performance by proximity to origin • Outliers clear • Useful for showing differing metric types e.g. bias and error • e.g. FB vs NMSE for DT&F



More – – – – and **less** – – – – stringent thresholds can be displayed

Analyses of results

Next presentation

Stop for questions

Aim	Which metrics*	Which datasets	Analyses details
Calculating metric ranges, with the aim of defining thresholds	*All normalised metrics*	*All*	*Analyse range of metric values from case studies; propose thresholds; and compare thresholds to literature values where appropriate*
Identifying the best metrics to quantify and analyse model error	Absolute and normalised error metrics	All	Consideration of metrics to quantify the magnitude of model error, and diagnosing possible sources of error
To demonstrate where metrics consistently identify the best and worst models in a model intercomparison	All	Datasets that include multiple models	Compare best and worst models within case study/trials, and across all case studies (for non-dimensional metrics); and assess if metrics consistently identify the best / worst models
Which metrics are suitable for assessing short- and long-term model performance	All	All	Review metric definitions to identify those that are only derived from period-average quantities as opposed to those that quantify point by point variations
Which metrics are suitable for datasets with few data points	All	'Small dataset, large set of models'	Identify particular quantities that are only valid for datasets with relatively large numbers of points, and exclude associated metrics

*except those already excluded

Model error: theory

- Useful to quantify model error
- Many commonly used error metrics are interlinked, for example:
  - RMSE is the square root of MSE
  - RMSE and ME both quantify **total error magnitude**, RMSE weighted towards higher conc, while ME is exact
  - MAE indicates period average model bias
  - CRMSE follows similar trends to RMSE

- Some model error metrics can be decomposed to investigate model performance and possible sources of error
- Mean Square Error (MSE):

Systematic component, the bias term, e.g. magnitude of emissions in the model: $\mathrm{MSEs} = \left(\overline{C_p} - \overline{C_o}\right)^2$

$$MSE = MSEs + MSEu$$

$$MSEu = MSE.cor + mMSE$$

Unsystematic component

Observed variance e.g. time variation of emissions in the model: $MSE.cor = \left(\sigma_p - r\sigma_o\right)^2$

Minimum achievable error which cannot be explained by systematic or correlated error e.g. model formulation: $mMSE = {\sigma_o}^2(1 - r^2)$

- Combined ranking metric (RANK) (N = number of observed/predicted pairs):

$$RANK = R^2 + \left(1 - \left|\frac{FB}{2}\right|\right) + FMS + \left(1 - \frac{KS}{100N}\right)$$

Correlation: error in the linear relationship e.g. temporal profiles of emissions

Bias: systematic differences in magnitude like over/under prediction e.g. inaccurate background conc

Exceedance agreement: degree of agreement from spatial differences between obs and mod

Similarity between **cumulative distributions of** the obs and mod

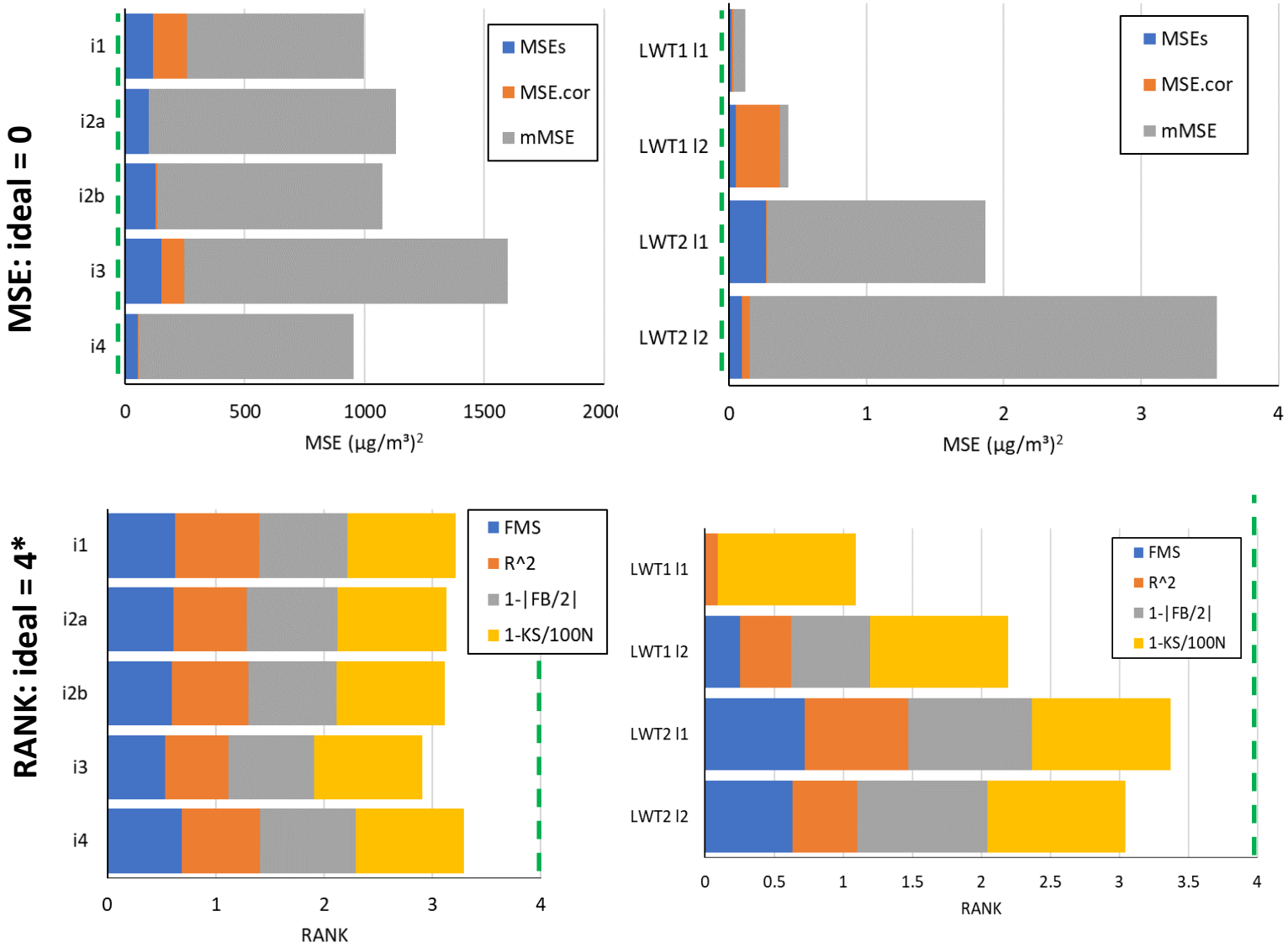
ADMLC CERC

Model error: results examples

- Idaho Falls:
  - MSE: large effect of mMSE indicates it may be difficult to improve model performance
  - RANK: no strong effect from the sources of error it decomposes
  - Worse (explainable) performance for model i3 for both metrics, but clearer signal from MSE
- Lee Wind Tunnel
  - MSE: again can see that in this case study, the studies with larger MSE are primarily due to the mMSE
  - RANK: Comparing models LWT1 l1 and l2, the higher RANK metric value is due to higher values of FMS, R and improved FB.
  - **MSE not directly comparable between trials (LWT1, LWT2) due to dimensionality**

Idaho Falls (*single trial, 5 models*)

Lee Wind Tunnel (*2 trials, 2 models*)



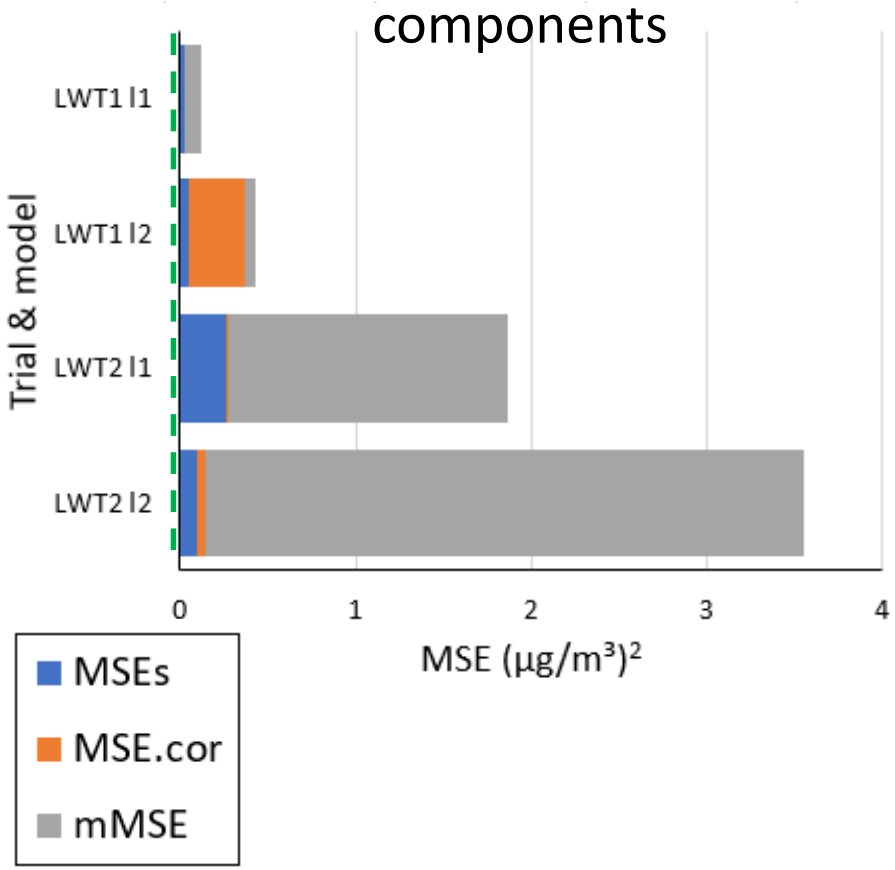
Ideal value

ADMLC CERC

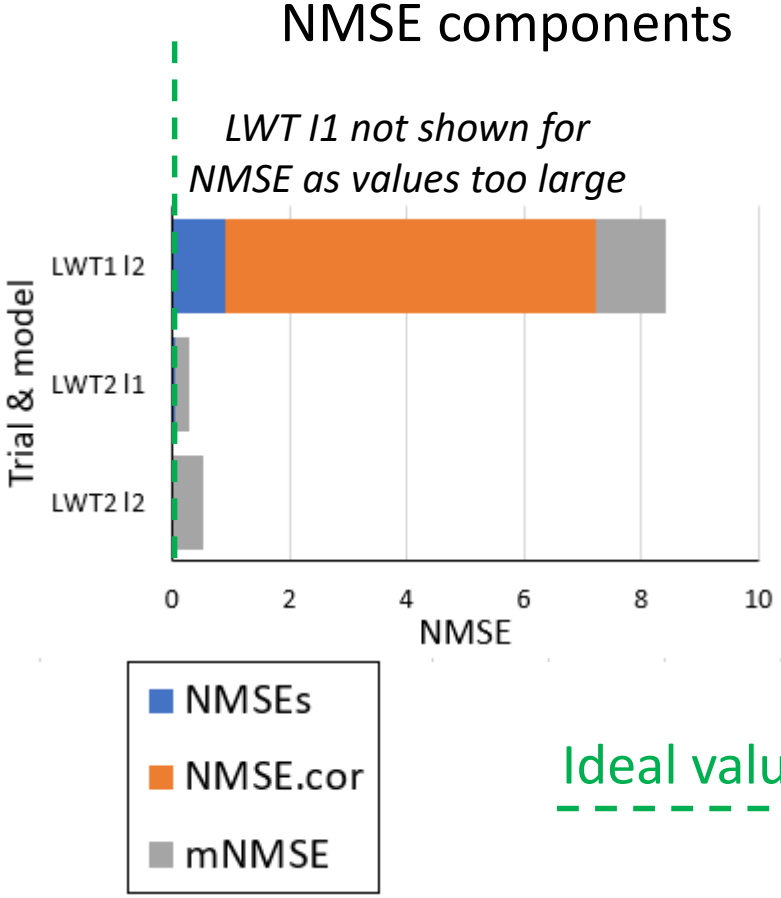
Model error: non-dimensional MSE -> nMSE

- Dimensional MSE useful for error diagnosis within a trial
- Non-dimensional MSE useful for comparisons across trials, & is consistent with RANK behaviour

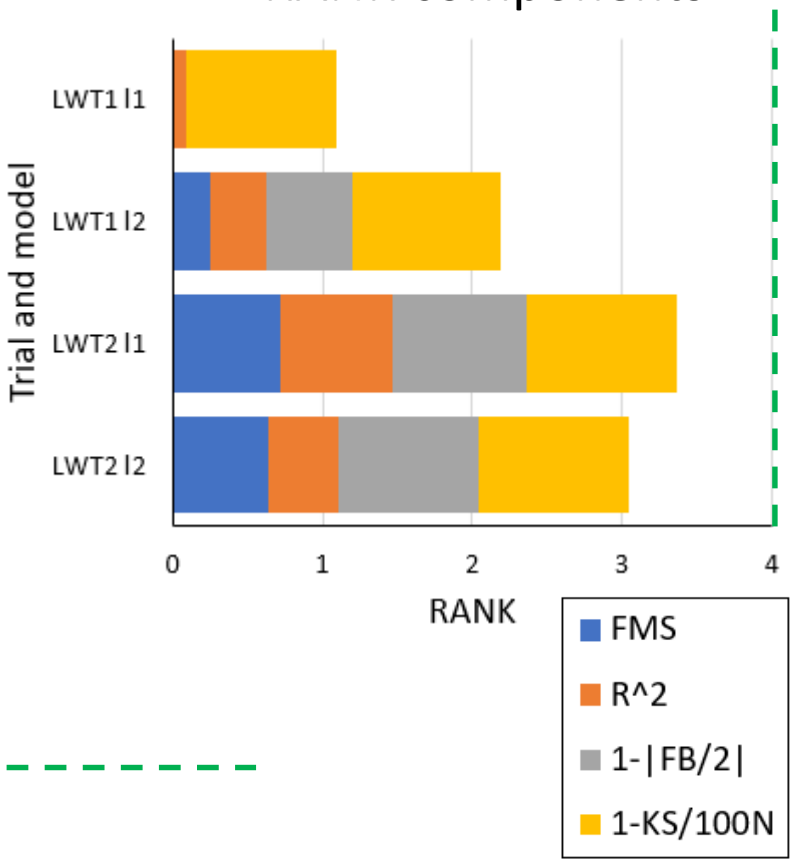
Dimensional MSE components



Non-dimensional NMSE components



Non-dimensional RANK components



Ideal value

Identifying the best and worst models: approach

- Within a trial:
  - Green/red filled cell indicates model closest to ideal/furthest from ideal

Difference (Normalised)			Normalised Mean Bias (NMB) *(non-dim)*	Fractional Bias (FB) *(non-dim)*	Mean Fractional Bias (MFB) *(non-dim)*	Normalised Mean Bias Factor (NMBF) *(non-dim)*
		Ideal	*0*	*0*	*0*	*0*
		Min	$-\infty$	-2	-2	$-\infty$
		Max	$+\infty$	$+2$	$+2$	$+\infty$
Category	**Trial**	**Model**	**NMB**	**FB**	**MFB**	**NMBF**
Large dataset, small set of models	K Q23	k1	0.016	0.016	N/A (0.043)	0.016
		k2	-0.585	-0.826	-0.809	-1.4
	K Q3	k1	-0.080	-0.083	-0.087	-0.086
		k2	-0.637	-0.936	-1	-1.8
	LWT s	l1	-1.0	**-2.0**	N/A (-0.242)	-8.93×10^{5}
		l2	1.5	0.863	N/A (0.672)	1.5
	LWT n	l1	-0.187	-0.206	N/A (-0.201)	-0.230
		l2	-0.112	-0.118	N/A (0.017)	-0.126
	C/A a1	c1a	4.74	1.41	N/A (-1.08)	4.74
		c1b	4.23	1.36	N/A (-0.58)	4.23
		c1c	3.03	1.20	N/A (-0.67)	3.03
	C/A a2	c1a	-0.500	-0.666	N/A (-1.2)	-0.998
		c1b	-0.376	-0.463	N/A (-0.623)	-0.602
		c1c	-0.543	-0.745	N/A (-0.830)	-1.2
	C/A c	c1a	0.425	0.350	N/A (-1.3)	0.425
		c1b	**-0.006**	**-0.006**	N/A (-0.821)	**-0.006**
		c1c	-0.010	-0.010	N/A (-0.323)	-0.010
	IF	i1	-0.306	-0.361	-0.178	-0.441
		i2a	-0.278	-0.323	-0.458	-0.386
		i2b	-0.319	-0.380	-0.33	-0.469
		i3	-0.346	-0.418	-0.365	-0.529
		i4	-0.207	-0.230	-0.127	-0.260

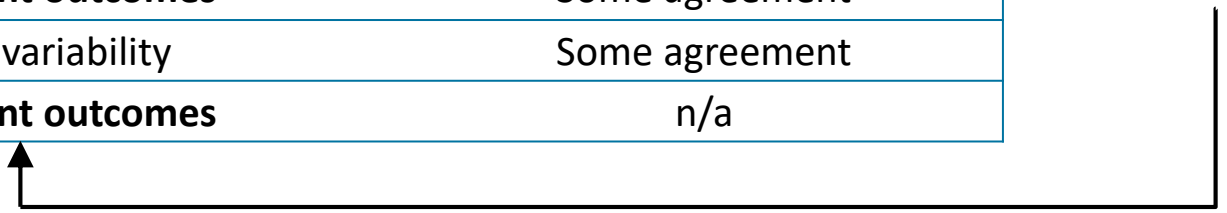
- Over all trials (non-dimensional metrics only):
  - **Green**/**red** bold text indicates model closest to ideal/ furthest from ideal

Identifying the best and worst models: outcomes

- Do metrics within particular metric types behave similarly?
  - *Broadly yes, although some are more consistent than others, and if there are many metrics within a category, there is more variation*
- How well do the metrics reflect 'overall performance, as quantified by the 'Summary / combination' metrics?
  - *There is some agreement for all metrics, but Error (Normalised) is the best reflection.*

Metric Type	Number of metrics	Variability of model performance within each metric type	Reflection of 'overall performance'?
Difference (Absolute)	1	n/a	Some agreement
Difference (Normalised)	3	**Consistent outcomes**	Some agreement
Error (Absolute)	9	Some variability	Some agreement
Error (Normalised)	12	Some variability	**Good agreement**
Fractional	3	Some variability	Some agreement
Linearity	3	Some variability	Some agreement
Area Exceedances	5	**Consistent outcomes**	Some agreement
Peak Concentrations	6	Some variability	Some agreement
Summary / combination	6	**Consistent outcomes**	n/a

'Some variability' implies care needs to be taken when choosing a metric

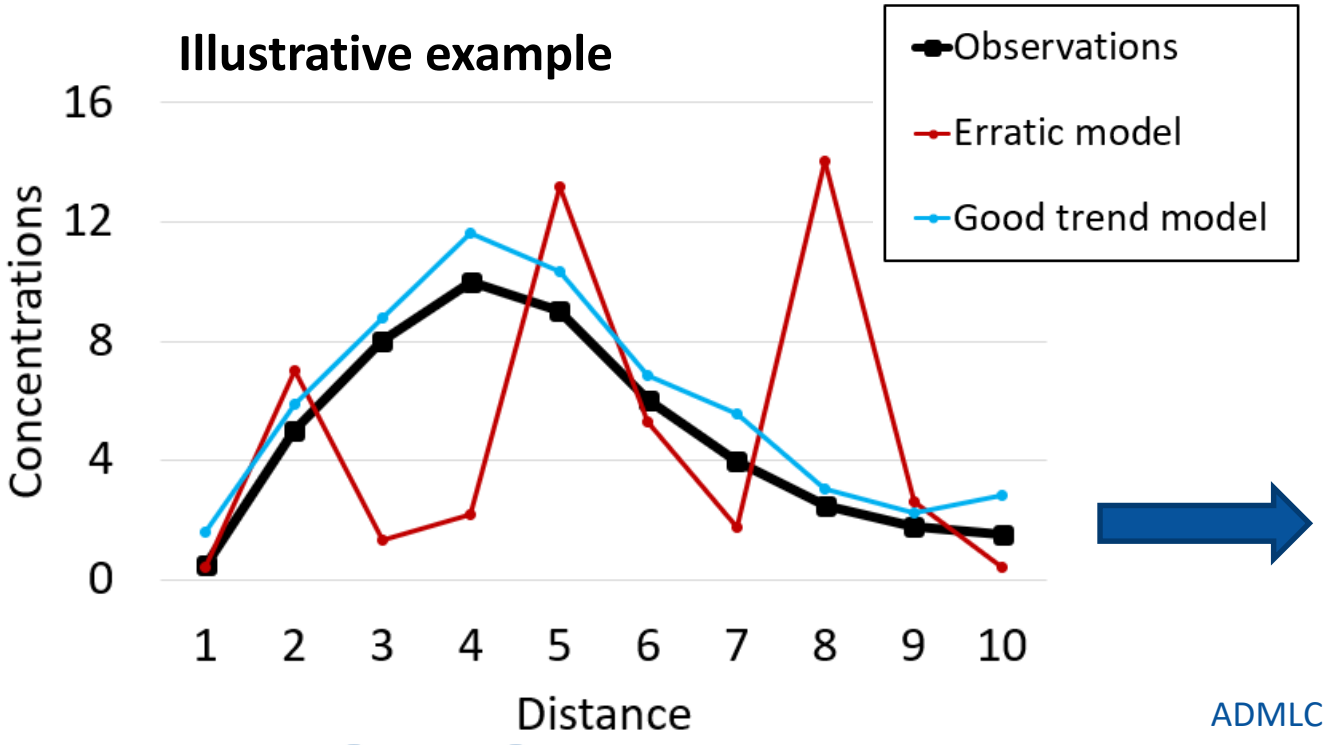


Statistical suitability of metrics: theory

- **Short- and long-term model performance**:
  - Long-term averaging compensates for short-term variability e.g. *mean concentrations*
  - Metrics have been evaluated based on whether they only contain period average terms or include terms which quantify time-resolved fluctuations

Only quantifies long-term performance	Quantifies short-term variability to some extent
e.g. Fractional Bias, FB	e.g. Normalised Mean Square Error, NMSE
$\mathrm{FB} = \frac{(\overline{C_p} - \overline{C_o})}{0.5(\overline{C_o} + \overline{C_p})}$	$\mathrm{NMSE} = \frac{\overline{(C_o - C_p)^2}}{\overline{C_o}\ \overline{C_p}}$
where: C_o is observed concentration C_p is modelled concentration	$\overline{C_o} = \sum C_o$

Illustrative example



Metric	Erratic model	Good trend model
Example long-term, FB	0	0.19
Example short-term, NMSE	1.15	0.04

Statistical suitability of metrics: theory

- **Few data points:**
  - ***FAIRMODE*** and ***Forecasting*** type metrics were identified early in the project as these are known to require a large amount of high quality data
  - Now consider in more detail, what statistical terms are unstable/unreliable for few data points: →
  - Based on statistical results (e.g. very little variation, extremely good performance, no value able to be calculated), the following metrics are likely unstable/unreliable:
    - NSD, FAC5, Rrank, RHC, KS

Terms considered not be robust for datasets with few data points:

- standard deviation
- variance
- covariance
- correlation
- regression slopes
- shape of the distribution
- particular extreme values: peak
- concentration exceedances

Statistical suitability of metrics: classification

- Based on these factors, metrics classified by suitability for quantifying:
  - short-term variability of a dataset or only suitable for assessing long-term model performance
  - datasets with few data points

Metric Type	Short-term variability	Only long-term model performance	Few data points
Difference (Absolute)		MB	MB
Difference (Normalised)		NMB, FB, NMBF	NMB, FB, NMBF
Error (Absolute)	ME, MSE, MSEu, MSE.cor, mMSE, RMSE, CRMSE	MAE, MSEs	ME, MAE, MSEs
Error (Normalised)	NAE, HH, NMSE, NMSEu, NMSE.cor, mNMSE, NSD, NRMSE, NMSD, Fs	NMSEs	NAE, HH, NMSEs
Fractional	FAC2, FACn, AE band		
Linearity	r, θ, rrank		
Area Exceedances	A_{FN}, A_{FP}, FMS, 1D MOE, 2D MOE		
Peak Concentrations	AU, AP, τ, ε, FMT	RHC	
FAIRMODE	MQI, MPI bias, MPI r, MPI sd	MQI annual, Spatial MPI r, Spatial MPI sd	
Forecasting	POD, SR, FBS, ACC, FAR, POFD, TS, GSS, ORSS, HSS, PSS, MPI FP, RANK2		
Summary / combination	IoA, CoE, KS, ρ2, λ, RANK		

Suitability of metrics for different applications: considerations

- Broad model application categories considered
- Characteristics associated of those types of model application datasets have been assumed in terms of:
  - Number, spatial and temporal distribution of measurements
  - Likelihood of model capturing the plume in time and space
- Assumptions:
  - Sufficient data points to allow calculation of metrics
  - Model type is suitable
  - Subset of measurements taken if required

Suitability of metrics for different applications: outcome

Field studies with intentionally released air pollutants

Unintentional release

Model application	Assumed dataset characteristics		Metric Type							
	Measurements	Likelihood of model capturing the plume in time and space	Difference	Error	Fractional	Linearity	Area Exceedances	Peak Concentrations	FAIRMODE	Forecasting
Short- and long-range tracer experiments / wind tunnel study	Large number, well distributed spatially; short time period	High	✓	✓	✓	✓	✓	✓		
Emergency response / accidental release	Low spatial coverage; short time period	Low	✓	✓	(✓)	(✓)		✓		
Routine monitoring (rural industrial sources)	Low spatial coverage; long time period	Low	✓	✓	(✓)	(✓)		✓		
Routine monitoring (urban sources)	Moderate spatial coverage; long time period	High	✓	✓	✓	✓	(✓)	✓	✓	
Urban air quality forecasting	Moderate or low spatial coverage; long time period	High	✓	✓	✓	✓			(✓)	✓

Recommendations (1 of 3): single model evaluation

Model users should:

- R.1: Generate graphics that provide a ‘sanity check’ on their model outputs
- R.2: Select a range of metrics to assess overall model performance
- R.3: Consider which metrics are most appropriate for their model dataset and application
- R.4: Use metrics and graphics to:
  - Describe how well their model performs
  - Diagnose sources of error
  - Assess overall model performance in relation to published thresholds / acceptance criteria
- R.5: Where model datasets include a large number of paired points that do not capture the plume, consider data filtering:
  - Include only data points that correspond to the higher measurements
  - Include data from a single wind sector (near-field, single source dispersion modelling)

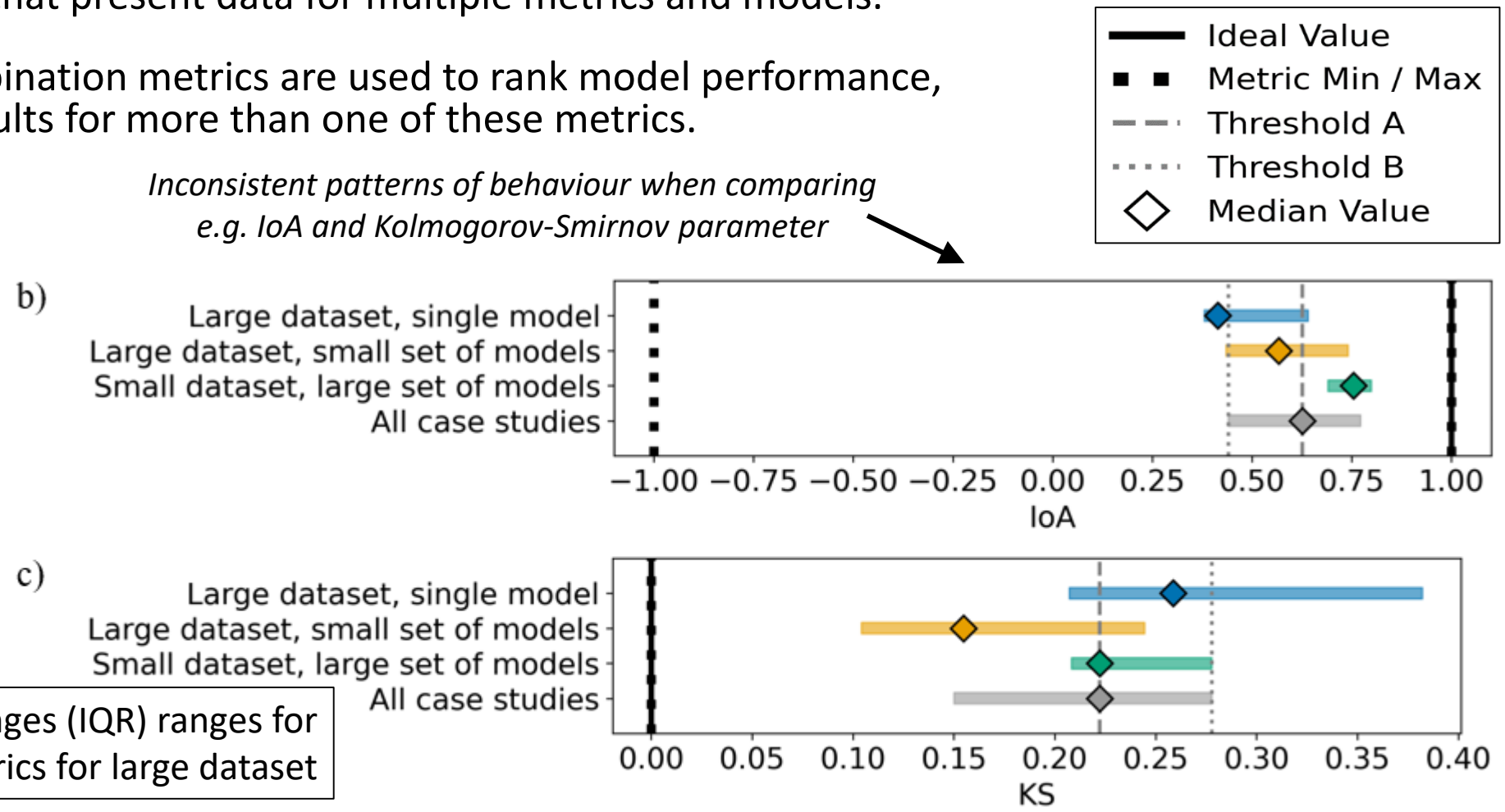
Further explanations are given in the report, including references to relevant sections

Recommendations (2 of 3): model inter-comparisons

Additional recommendations for model inter-comparisons:

- R.6: Generate graphics that present data for multiple metrics and models.
- R.7: If summary & combination metrics are used to rank model performance, consider presenting results for more than one of these metrics.

Inconsistent patterns of behaviour when comparing e.g. IoA and Kolmogorov-Smirnov parameter



Further explanations are given in the report, including references to relevant sections

Interquartile ranges (IQR) ranges for Summary/Combination metrics for large dataset

Recommendations (3 of 3): further work

Additional project recommendations:

- R.8: Metric thresholds calculated should be reviewed by the AQ modelling community in order for them to become acceptance criteria e.g. dedicated workshop and / or peer-reviewed journal article
  - **Here we are!**
- R.9: Release an updated version of the Model Evaluation Toolkit that includes the reliable metrics, making it publicly available to the AQ modelling community
  - **Completed! See afternoon presentations**
- R.10: Further work should include assessing the value of accounting for statistical significance in relation to model ranking within model inter-comparison studies
- R.11: Further work could include development of methods to allow for measurement uncertainty in metric calculations and / or model evaluation graphics

Any questions or comments?