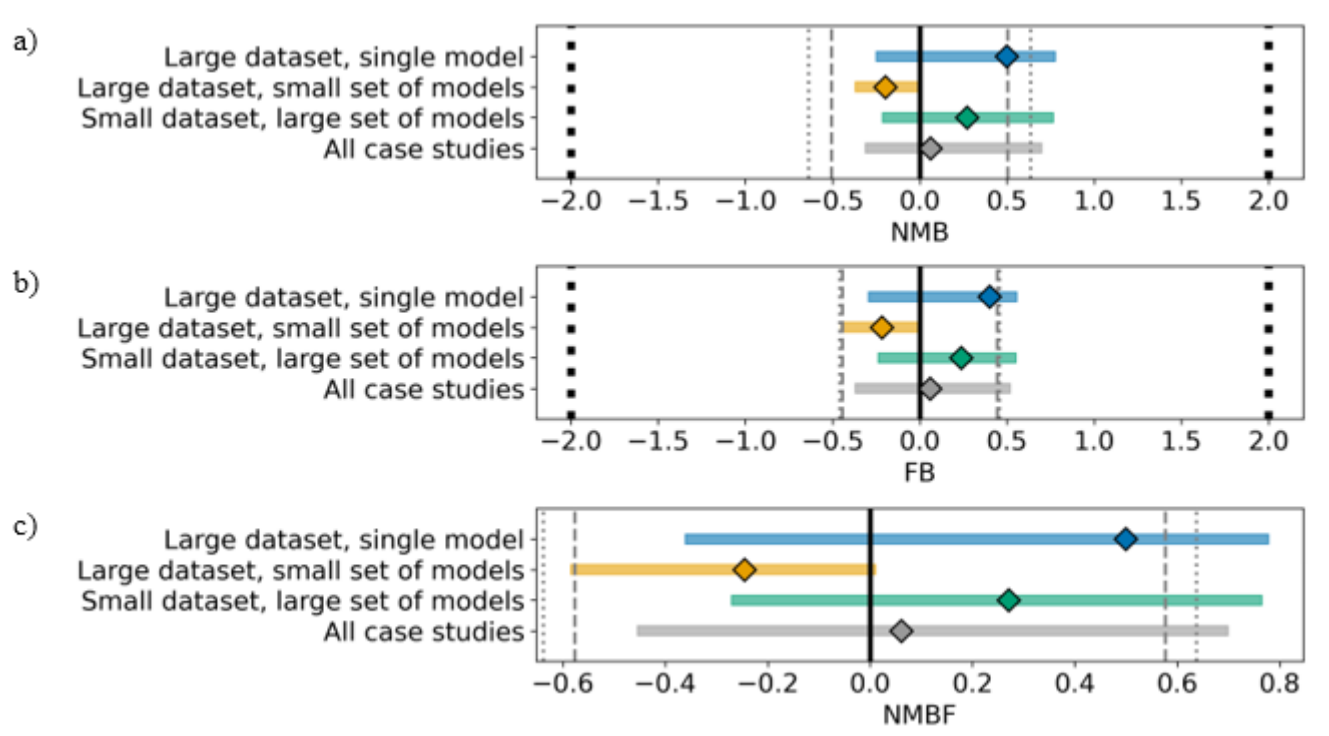
Threshold method and results



Victoria Hamilton

Model Evaluation Workshop: Thresholds & Toolkit

UKHSA Training Centre

11th June 2026

Analyses of results

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

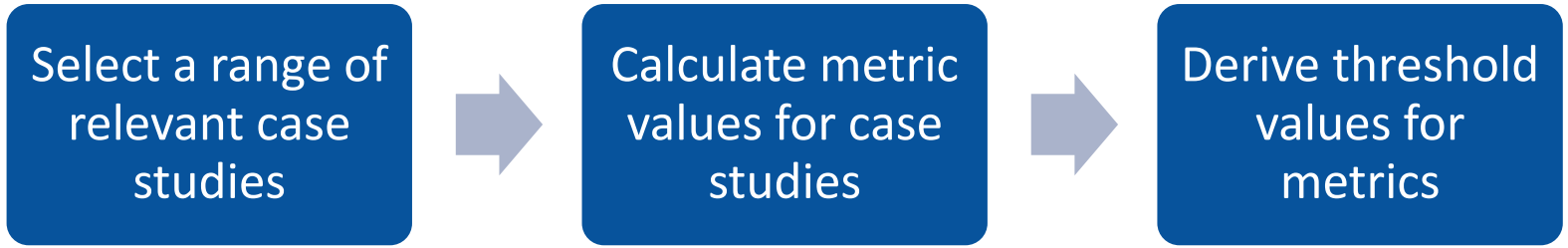
Which metrics does it make sense to propose thresholds for?

- Can only calculate thresholds for **non-dimensional** metrics
- *FAIRMODE* & *Forecasting* are non-dimensional but insufficient datasets and some FAIRMODE metrics have thresholds
- Thresholds were also proposed for **error components** (shown in *italics*) – should only MSE and NMSE should be assessed?

Metric type	DIMENSIONAL	NON-DIMENSIONAL	
	Metric(s)	Metric(s)	Threshold?
Difference (Absolute)	MB		
Difference (Normalised)		NMB, FB, NMBF	Y
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	Y
Fractional		FAC2, FACn, AE band	Y
Linearity		r, rrank, ϑ	Y
Area Exceedances	A_{FN}, A_{FP}	FMS, 1D MOE, 2D MOE (X_1, X_2)	Y
Peak Concentrations	τ, ε, FMT	AU, AP, FMT	Y
FAIRMODE		MQI, MQI annual, MPI bias, MPI r, MPI sd, Spatial MPI r, Spatial MPI sd	N
Forecasting		POD, SR, FBS, ACC, FAR, POFD, TS, GSS, ORSS, HSS, PSS, MPI FP	N
Summary / combination		IoA, CoE, KS, RANK2, MSI ($p2$), MMSI (λ)	Y

Examples of published model performance criteria

- "Acceptance criteria": metric thresholds which have already been accepted for general use in model evaluation
- To determine thresholds typically:



- Chang & Hanna (2004, 2012):
  - "Acceptance criteria"; general atmospheric dispersion model evaluation using field/tracer datasets; widely applied across accidental and routine dispersion studies
- Ivings (2013):
  - "Statistical Performance Measures"; developed for US LNG vapour dispersion regulatory assessment; based on liquefied natural gas spill and dense gas datasets
- FAIRMODE (Thunis and Cuvelier, 2022):
  - "Model Performance Metrics"; regulatory air quality model evaluation against monitoring data; commonly applied for long-term compliance assessment of ambient air quality standards

Metric thresholds: Chang & Hanna acceptance criteria

- Chang & Hanna reviewed the statistical performance of widely used and evaluated models (2004, 2012)
- They then proposed what values "good" models typically give, which allowed them to empirically determine acceptance criteria (shown in the table below)

Metric	Data	Acceptance Criteria (C&H)	
		Rural	Urban
FB	Arc-max	$\lvert FB \rvert \lesssim 0.3$	$\lvert FB \rvert \lesssim 0.67$
NMSE	Arc-max	$NMSE \lesssim 3$	$NMSE \lesssim 6$
FAC2	Arc-max	$FAC2 \gtrsim 0.5$	$FAC2 \gtrsim 0.3$
NAD/NAE	Paired in space	$NAD \lesssim 0.3$ $NAE \lesssim 0.6$	$NAD \lesssim 0.5$ $NAE \lesssim 1.0$
Overall criterion:		**Criteria should be satisfied at least half the time, in at least half of the field experiments**	

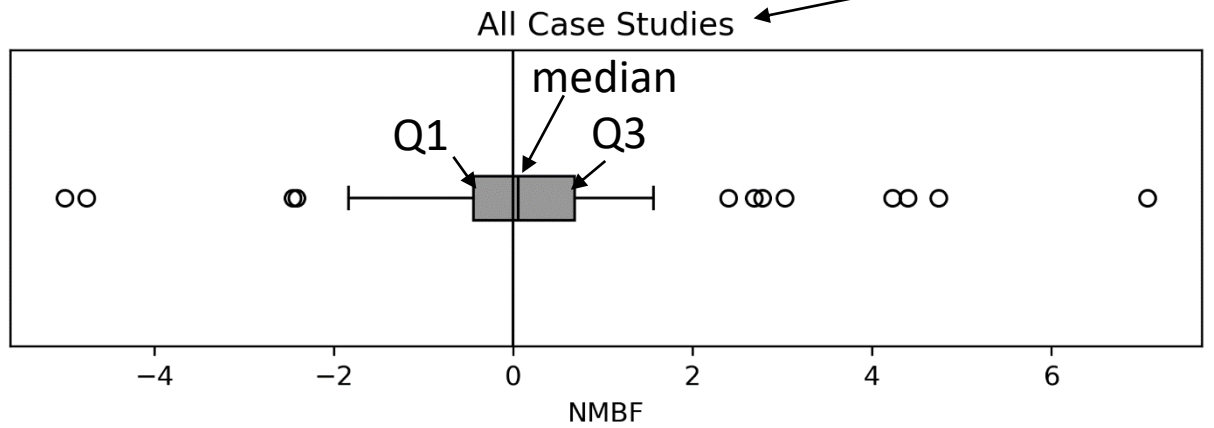
Metric thresholds: methodology for thresholds

- Aim to extend the Chang & Hanna method
- Key differences:
  - **No restrictions**: on data type (e.g. arc-max) or study setting (e.g. urban/rural)
  - Assume models represent a **range of performance** (rather than considering which are “good”)
  - We calculate the metrics for all the case studies, which gives us a range of metric values
    - We use this **range** to indicate “good” and “less-good” model performance

Box plots: basics

- **Key components:**
  - **Box:** Interquartile range (IQR)
    - IQR = Q3 – Q1
      - Q1 = 25th percentile
      - Q3 = 75th percentile
  - **Whiskers:** 1.5×IQR from Q1/Q3
  - **Median (Q2):** Line inside the box
  - **Outliers:** Points beyond whiskers

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	K1	k1	0.016	0.016	N/A (0.043)	0.016
		k2	-0.585	-0.826	-0.809	-1.4
	K2	k1	-0.080	-0.083	-0.087	-0.086
		k2	-0.637	-0.936	-1	-1.8
	LWT1	l1	-1.0	-2.0	N/A (-0.242)	-8.93×10^{5}
		l2	1.5	0.863	N/A (0.672)	1.5
	LWT2	l1	-0.187	-0.206	N/A (-0.201)	-0.230
		l2	-0.112	-0.118	N/A (0.017)	-0.126
	C/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/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/A3	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	-0.330	-0.469

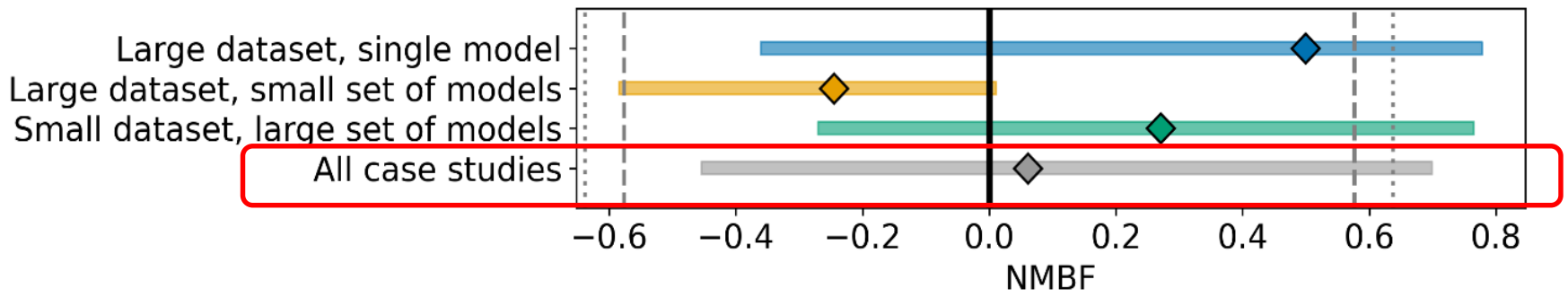


Box plots: report method

- Method:
  - Based around **interquartile ranges (IQR)** from our range of model performances across the case studies.
  - Propose two thresholds where possible: more stringent (**A**) & less stringent (**B**)
  - Exact method is dependent on metric properties e.g. upper/lower bounded, symmetric/non-symmetric

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	K1	k1	0.016	0.016	N/A (0.043)	0.016
		k2	-0.585	-0.826	-0.809	-1.4
	K2	k1	-0.080	-0.083	-0.087	-0.086
		k2	-0.637	-0.936	-1	-1.8
	LWT1	l1	-1.0	-2.0	N/A (-0.242)	-8.93×10^{5}
		l2	1.5	0.863	N/A (0.672)	1.5
	LWT2	l1	-0.187	-0.206	N/A (-0.201)	-0.230
		l2	-0.112	-0.118	N/A (0.017)	-0.126
	C/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/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/A3	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	-0.330	-0.469

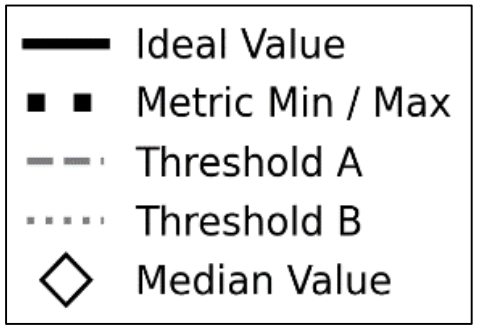
- Interpreting the box plot:
  - **'All case studies'** (grey box): range of model performance used to calculate thresholds
  - **Thresholds** shown as grey dashed line
  - **3 dataset categories**: data set size, number of models size shown as reference e.g. check for abnormal skew
  - **IQR box** and **median** (diamond): statistics used for calculating thresholds
  - **Ideal value** (black line)
  - **Min/Max** (black dashed lines) (not shown here as NMBF min/max = ±∞)

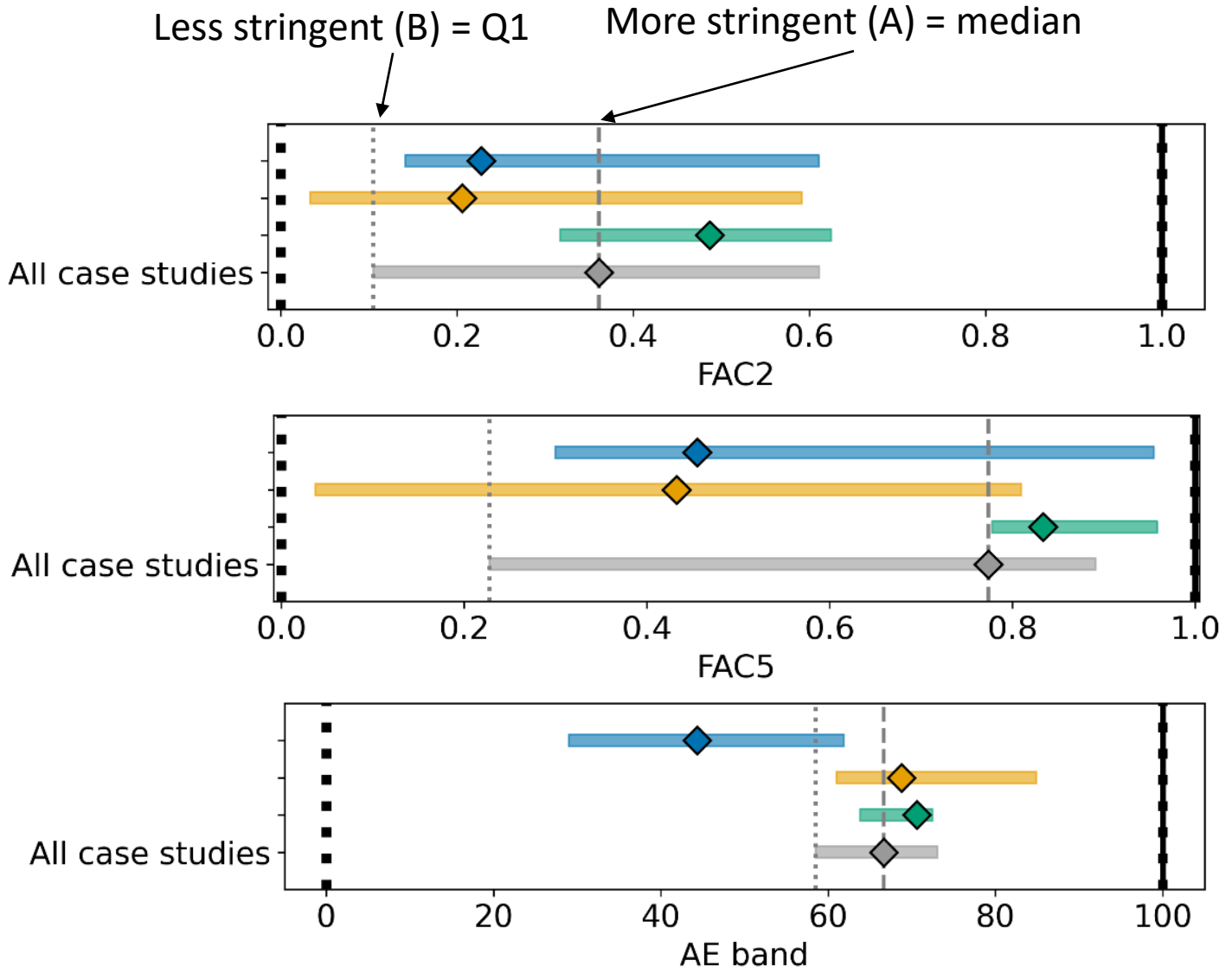


Metric thresholds: ***Fractional*** results example

Metric	Ideal	Min	Max
FAC2	1	0	1
FAC5	1	0	1
AE band	100	0	100

Asymmetrical, where max is the ideal value: define lower threshold

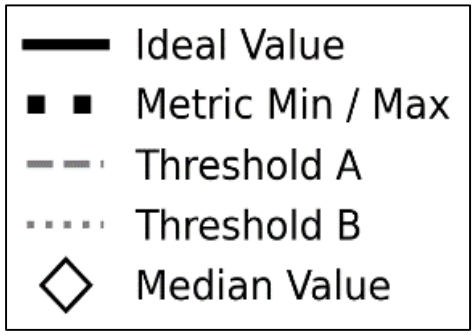




Metric thresholds: ***Difference (normalised)*** results example

Metric	Ideal	Min	Max
NMB	0	−∞	+∞
FB	0	−2	+2
NMBF	0	−∞	+∞

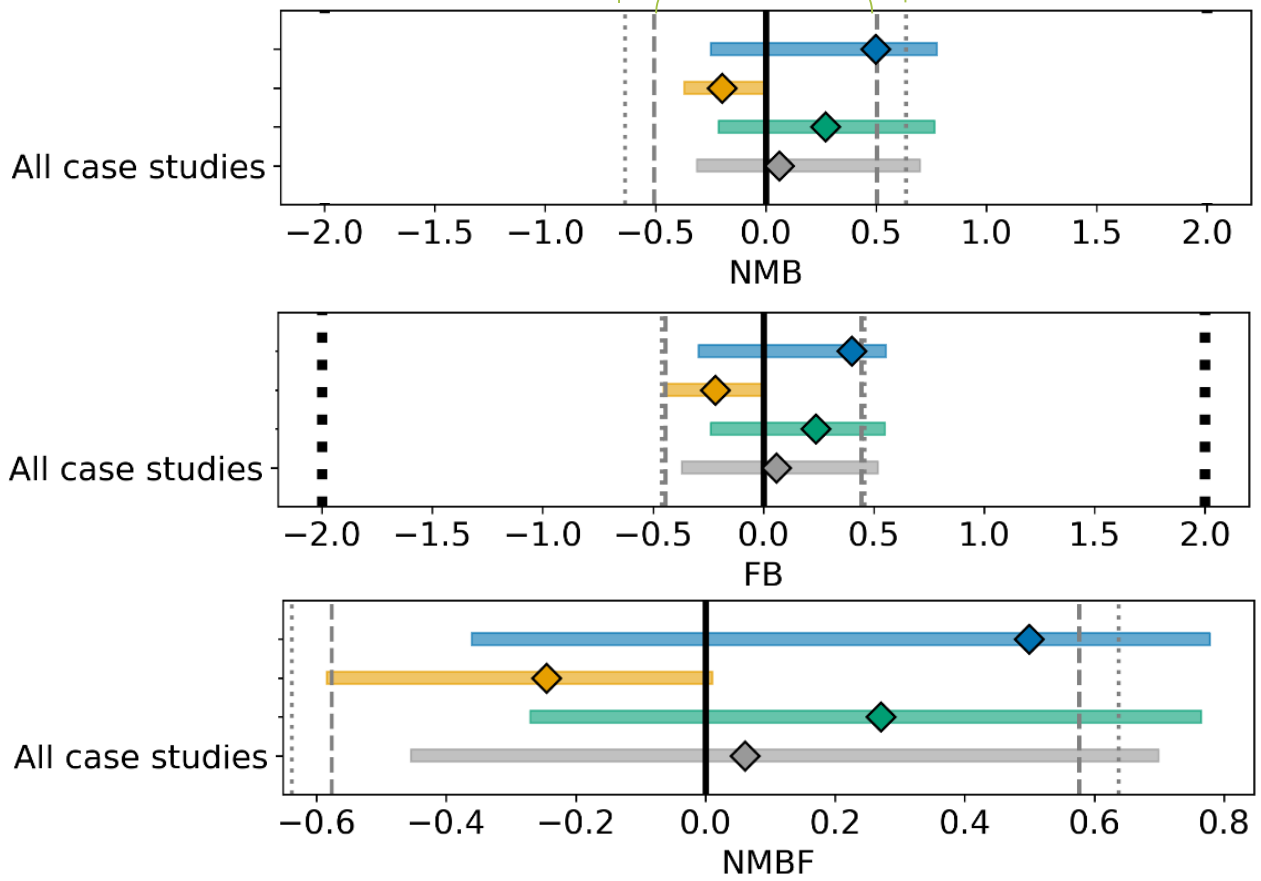
Symmetrical about zero, where zero is the ideal value: define symmetric threshold



e.g. NMB (Q3 – median) > (median – Q1)

Less stringent (B) = define half-width; applied symmetrically about ideal

More stringent (A) = IQR moved to be symmetrical about ideal



ADMLC CERC

C&H acceptance criteria dataset pre-requisites

- Which of our case studies are applicable to the C&H acceptance criteria?
- e.g. no Urban arc-max datasets for calculating FB, NMSE, FAC2

Metric	Acceptance Criteria (Chang & Hanna)		
	Dataset type	Criteria	
		Rural	Urban
|FB|	Arc-max	More stringent	Less stringent
NMSE			
FAC2			
NAD/NAE	Paired in space		

Study	Data	Location	Suitable criteria
K	Arc-max	Rural	FB, NMSE, FAC2
LWT	Arc max	Rural	FB, NMSE, FAC2
WM-Air	Cont. hourly	Urban	NAE
AT	Cont. hourly	Urban	NAE
C/A	3 hourly/ 24-hour avg.	Rural	NAE
DT/F*	Avg. arc max	Rural	FB, NMSE
IF	15-min avg.	Rural	NAE
JRII	Avg. arc-max	Rural	FB, NMSE, FAC2

*FAC2 not calculated due to small dataset size

Comparison of thresholds with C&H acceptance criteria

- Summary of **Chang & Hanna's 'Acceptance Criteria'** for suitable case studies, compared with **our calculated 'Threshold Scenarios'** from scenarios A (more stringent) and B (less stringent). Results are fairly comparable:

Metric	Acceptance Criteria (Chang & Hanna)			Threshold Scenarios (ADMLC Report)		
	Dataset type	Criteria		Dataset type	Threshold	
		Rural	Urban		A	B
|FB|	Arc-max	$\lesssim 0.3$	$\lesssim 0.67$	All	$\lesssim 0.44$	$\lesssim 0.46$
NMSE		$\lesssim 3$	$\lesssim 6$		$\lesssim 2.5$	$\lesssim 16.8$
FAC2		$\gtrsim 0.5$	$\gtrsim 0.3$		$\gtrsim 0.36$	$\gtrsim 0.11$
NAE	Paired in space	$\lesssim 0.6$	$\lesssim 1.0$		$\lesssim 0.73$	$\lesssim 1.33$

- **FB:** threshold scenarios A & B are both approximately in-between acceptance criteria for rural and urban
- **NMSE**: scenario A slightly more stringent than acceptance criteria, while scenario B is much less stringent
- **FAC2**: scenario A is close to the Urban acceptance criteria
- **NAE:** scenario A is less stringent than the Urban acceptance criteria

Calculating pass rates

- '**Pass rate**' = % of trials which satisfy the criteria/threshold

- Example shows pass/fail for FB for:
  - Rural acceptance criteria (C&H)
  - Threshold A (more stringent) (ADMLC report)
  - Threshold B (less stringent) (ADMLC report)

		Acceptance Criteria (Chang & Hanna)	Threshold Scenarios (ADMLC Report)	
	Metric	Rural Criteria	A Threshold	B Threshold
Trial	$\lvert FB \rvert$	$\lesssim 0.3$	$\lesssim 0.44$	$\lesssim 0.46$
K1 k1	0.016	✓	✓	✓
K1 k2	0.826	✗	✗	✗
K2 k1	0.083	✓	✓	✓
K2 k2	0.936	✗	✗	✗
LWT1 l1	2.0	✗	✗	✗
LWT1 l2	0.863	✗	✗	✗
LWT2 l1	0.206	✓	✓	✓
LWT2 l2	0.118	✓	✓	✓
j1	0.477	✗	✗	✗
j2	0.205	✗	✗	✗
j3	0.229	✓	✓	✓
j4	0.826	✗	✗	✗
j5	1.3	✗	✗	✗
j6	0.208	✓	✓	✓
j7	0.269	✓	✓	✓
j8	0.234	✓	✓	✓
j9	0.217	✓	✓	✓
j11	0.877	✗	✗	✗
j12	0.425	✗	✓	✓

Comparison of thresholds and pass rates with C&H

- Comparison of '**pass rates**': % of trials over all applicable case studies which satisfy the criteria/threshold

Metric	Acceptance Criteria (Chang & Hanna)				Threshold Scenarios (ADMLC Report)			
	Rural		Urban		A		B	
	Criteria	Pass rate (%)	Criteria	Pass rate (%)	Threshold	Pass rate (%)	Threshold	Pass rate (%)
|FB|	$\lesssim 0.3$	58.1	$\lesssim 0.67$	-	$\lesssim 0.44$	50.7	$\lesssim 0.46$	52.2
NMSE	$\lesssim 3$	69.8	$\lesssim 6$	-	$\lesssim 2.5$	50.7	$\lesssim 16.8$	74.6
FAC2	$\gtrsim 0.5$	50	$\gtrsim 0.3$	-	$\gtrsim 0.36$	50	$\gtrsim 0.11$	75
NAE	$\lesssim 0.6$	37.3	$\lesssim 1.0$	58.2	$\lesssim 0.73$	50.7	$\lesssim 1.33$	74.6

- **FB**: comparable pass rate
- **NMSE:** B pass rate is comparable
- **FAC2:** pass rate is comparable
- **NAE**: pass rate comparable to urban acceptance criteria rate

All calculated thresholds

- See the handout all calculated threshold results and some useful details

Metric type | Metric names | Thresholds | Metric info

FULL LIST OF METRICS FOR WHICH THRESHOLDS WERE CALCULATED, INCLUDING THRESHOLD CRITERIA, MIN/MAX AND IDEAL VALUES

Metric Type	Metric Short-Name	Metric Long-Name	Threshold Criteria				Ideal	Min	Max
			A		B				
			Lower	Upper	Lower	Upper			
Difference (Normalised)	NMB	Normalised Mean Bias	-0.51	0.51	-0.64	0.64	0	−∞	+∞
	FB	Fractional Bias	-0.44	0.44	-0.46	0.46	0	-2	+2
	NMBF	Normalised Mean Bias Factor	-0.58	0.58	-0.64	0.64	0	−∞	+∞
Error (Normalised)	NMSE	Normalised Mean Square Error	-	2.5	-	16.8	0	0	+∞
	NMSEs	Component of NMSE due to systematic errors	-	0.2	-	1.01	0	0	+∞
	NMSEu	Component of NMSE due to random errors	-	1.9	-	15.6	0	0	+∞
	NMSE.cor	Variance correlation component of NMSE	-	1.5	-	5.9	0	0	+∞
	mNMSE	Minimum achievable NMSE	-	0.73	-	1.92	0	0	+∞
	NAE	Normalised Absolute Error	-	0.73	-	1.334	0	0	+∞
	HH	Symmetrically normalised RMSE	-	0.86	-	2.24	0	0	+∞
	NSD	Normalised Standard Deviation	0.55	1.62	-	-	1	0	+∞
	NRMSE	Normalised Root Mean Square Error	-	1.6	-	4.1	0	0	+∞
	NCRMSE	Normalised Centred Root Mean Square Error	-	0.934	-	1.19	0	0	+∞
	NMSD	Normalised Mean Standard Deviation	-0.45	0.62	-	-	0	-1	+∞
	Fs	Fractional Standard deviation	-0.61	0.61	-0.69	0.69	0	-2	2
Fractional	FAC2	Fraction of model predictions within a FACtor of 2	0.42	-	0.15	-	1	0	1
	FAC5	Fraction of model predictions within a FACtor of 5	0.77	-	0.23	-	1	0	1
	AE band	Absolute error band	66.7	-	58.5	-	100	0	100
Linearity	r	Correlation coefficient	0.623	-	0.299	-	1	-1	1
	rrank	Spearman correlation coefficient (Ranked correlation)	0.733	-	0.405	-	1	-1	1
	θ / theta	Linear least-squares slope parameter	0.497	1.503	0.356	1.644	1	−∞	+∞
Area Exceedances	FMS	Figure of Merit in Space	0.11	-	0.07	-	1	0	1
	1D MOE	User-oriented 1-Dimensional measure of effectiveness	0.05	-	0.04	-	1	0	1
	2D MOE X_1	User-oriented 2-Dimensional measure of effectiveness	0.17	-	0.14	-	(1, 1)	(0, 0)	(1, 1)
	2D MOE X_2	User-oriented 2-Dimensional measure of effectiveness	0.29	-	0.16	-	(1, 1)	(0, 0)	(1, 1)
Peak Concentrations	AU	Accuracy of Unpaired peak	-0.54	0.96	-	-	0	-1	+∞
	AP	Accuracy of Paired peak	-0.91	0.37	-	-	0	-1	+∞
	FMT	Figure of Merit in Time	44.2	-	18.5	-	100	0	100
Summary / combination	IoA	Index Of Agreement	0.626	-	0.441	-	1	-1	1
	COE	Coefficient Of Efficiency	0.252	-	-0.118	-	1	−∞	1
	KS	Kolmogorov-Smirnov parameter	-	0.222	-	0.278	0	0	N
	RANK	Combined Ranking metric	2.16	-	1.75	-	4	0	4
	ρ2 / MSI	Mielke square index	0.464	-	0.207	-	1	-1	1
	λ / MMSI	Modified Mielke index	0.464	-	0.207	-	1	0	1

Other analyses

- From the thresholds we also considered:
  - Overall % best/worst performing models for a particular trial and how that compares within metric type and across metric types (*Section 6.4* & *Section 7*) (discussed in previous presentation)
  - Overall pass rate for each trial averaged by metric type (*Section 7)*

Category	Trial	Model	Fraction assigned as best (B) or worst (W) performing model for each trial by metric type:																			
			Difference Absolute (/1)		*Difference Normalised* (/3)		*Error (Absolute)* (/9)		*Error Normalised* (/12)		*Fractional* (/3)		*Linearity* (/3)		*Area Exceedances* (/5)		*Peak* (/6)		*Summary / combination* (/6)		Overall (%)	
			B	W	B	W	B	W	B	W	B	W	B	W	B	W	B	W	B	W	B	W
	K1	k1	1	0	3	0	8	1	11	1	3	0	2	1	-	-	3	3	5	0	87.2	12.8
		k2	0	1	0	3	1	8	1	11	0	3	1	2	-	-	3	3	0	5	12.8	87.2
	K2	k1	1	0	3	0	6	3	9	3	3	0	2	1	-	-	3	3	5	0	82.3	17.7
		k2	0	1	0	3	3	6	3	9	0	3	1	2	-	-	3	3	0	5	17.7	82.3
	LWT1	l1	1	0	1	2	8	1	3	9	1	2	0	3	1	4	3	3	2	4	42.7	57.3
		l2	0	1	2	1	1	8	9	3	2	1	3	0	4	1	3	3	4	2	57.3	42.7
	LWT2	l1	0	1	0	3	7	2	11	1	2	1	3	0	5	0	6	0	5	1	68.8	31.2
		l2	1	0	3	0	2	7	1	11	1	2	0	3	0	5	0	6	1	5	31.2	68.8

Trial	Model	Overall	
		Pass rate (%)	Fraction Best (%)
K1	k1	76.7	87.2
	k2	20.8	12.8
K2	k1	68.6	82.3
	k2	26.4	17.7
LWT1	l1	9.3	42.7
	l2	15.5	57.3
LWT2	l1	90.5	68.8
	l2	95.2	31.2
C/A1	c1a	4.8	9.6
	c1b	7.6	17.2
	c1c	7.6	73.2
C/A2	c1a	15.5	5.6
	c1b	26.2	66.9
	c1c	8.3	27.5
C/A3	c1a	23.1	8.3
	c1b	26.7	31.5
	c1c	38.6	60.1
	i1	94.1	15.2
	i2a	89.3	0

Any questions or comments?