Evaluating urban climate models

DR HELEN CLAIRE WARD

UNIVERSITY OF INNSBRUCK

OVERVIEW

Introduction

- Why is evaluation important and what does it tell us?
- Types of urban climate models
- Types of reference datasets

General considerations

- Quantities
- Processes and scales
- Assumptions and uncertainties
- Compatibility and independence

Temporal considerations

- Evaluation period and averaging interval

Spatial considerations

- Spatial representativeness of reference dataset
- Matching grid points horizontally and vertically

Examples

- Various examples with different models and reference datasets

Good practice

- Important checks
- Use of statistics
- Designing figures
- Limitations of the evaluation

WHY IS EVALUATION IMPORTANT?

Urban climate models are widely (and increasingly) used to provide information and aid decision-making, often with direct implications for -

THE ECONOMY

THE ENVIRONMENT

HUMAN HEALTH

It is therefore **critical** that model output is sufficiently accurate and reliable for the application

Decision-making is often based around possible scenarios, so we need to ensure that the model is

"right for the right reasons"

The challenge is to both assess and understand model performance

WHY IS EVALUATION IMPORTANT?

Without careful model setup, interpretation and evaluation, we risk

- unawareness of limitations
- poor understanding of interdependencies and potential inadvertent effects
- provision of wrong information
- misinformed decisions, leading to undesirable or even dangerous situations

https://www.deeproot.com/blog/blog-entries/the-best-street-trees-reader-edition-results/

Some examples of positive and negative effects:

- street trees may improve the thermal environment but can trap pollutants close to the surface
- increasing albedo may reduce energy input to the surface but can negatively effect pedestrian comfort
- urban cooling measures may reduce urban temperatures but can reduce air quality

Without understanding the model, we cannot hope to develop/improve performance or advance current knowledge

WHAT DOES EVALUATION TELL US?

Model evaluation actually only tells about how model output compares to the reference dataset under the conditions tested

The goal is (usually) to design the evaluation so that we learn more than this specific case and can build up confidence in the model more generally

This can be achieved through

- evaluation over a long time-period or covering a range of conditions
- evaluation at multiple different locations
- consideration of multiple variables in order to assess processes and interactions
- sensitivity studies to check for robustness and quantify dependencies
- thorough understanding of how the model works

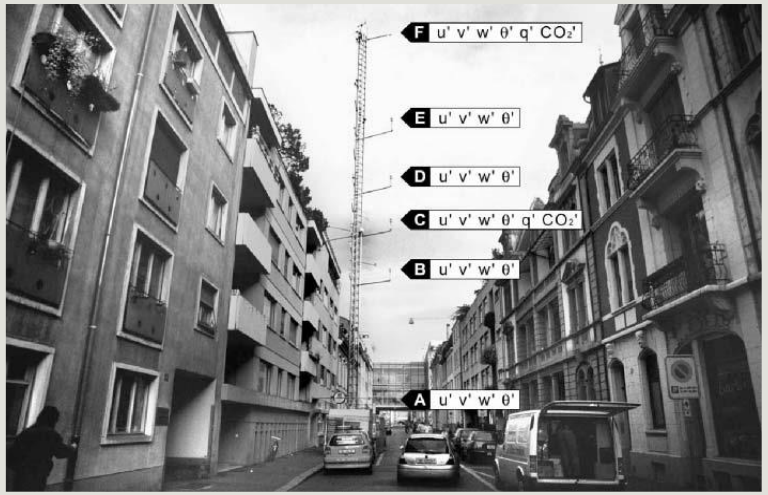
- weather and climate
- geographical setting
- urban characteristics
- socio-economic factors

The design of the evaluation depends on the purpose/application, the type of model, the computational resources available and the reference datasets available

TYPES OF REFERENCE DATASETS

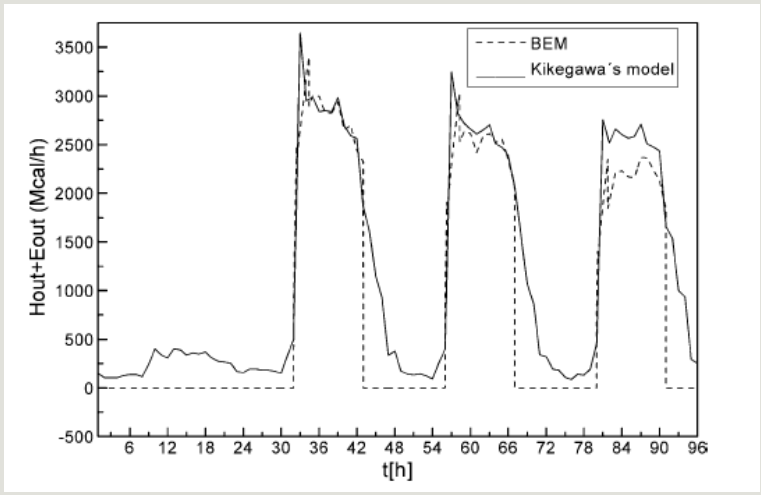
A model (or model version) can be evaluated against a reference dataset, such as

- real-world observations
- experimental data (e.g. wind-tunnel or water-tank)
- output from other models (e.g. different versions of the same model; different types of model)



Rotach et al. (2005) TAC

Gronemeier et al. (2021) GMD



Salamanca et al. (2010) TAC

TYPES OF URBAN CLIMATE MODELS

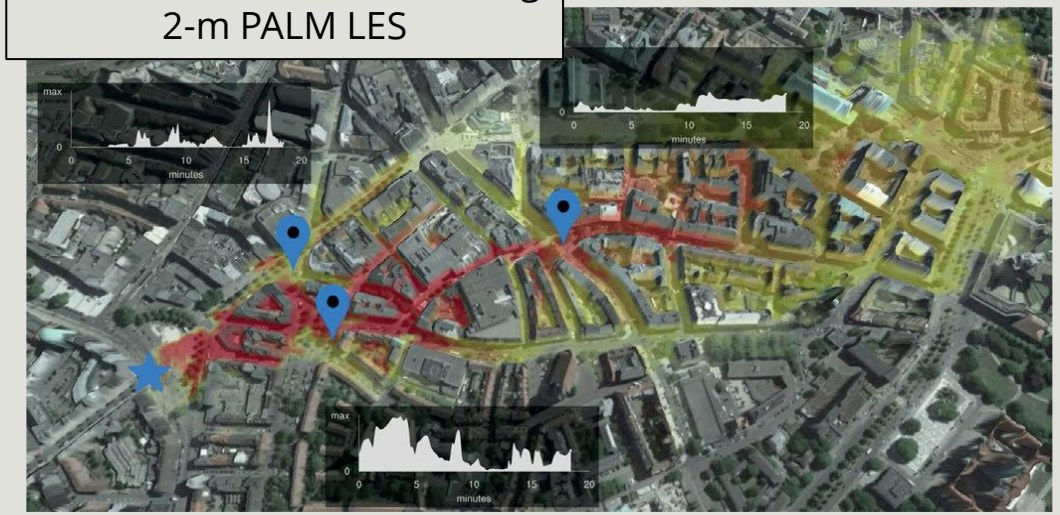
- There is a broad range of models relevant to urban climate

simple	highly complex
single process/quantity	multiple processes/quantities
single point	regional/global
fine scale	coarse scale
1D	3D
general	urban

- Focus here on micro-scale to meso-scale models and their urban capabilities (e.g. PALM, WRF…) and urban land-surface schemes

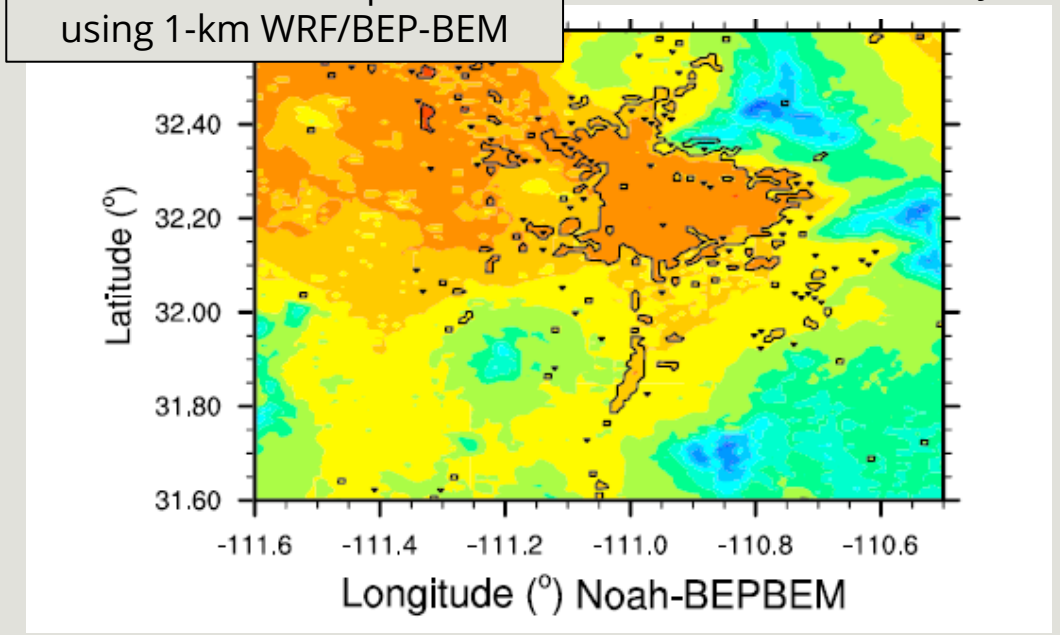
Pollutant concentration using 2-m PALM LES

Maronga et al. (2019) MetZet



Mean surface temperature using 1-km WRF/BEP-BEM

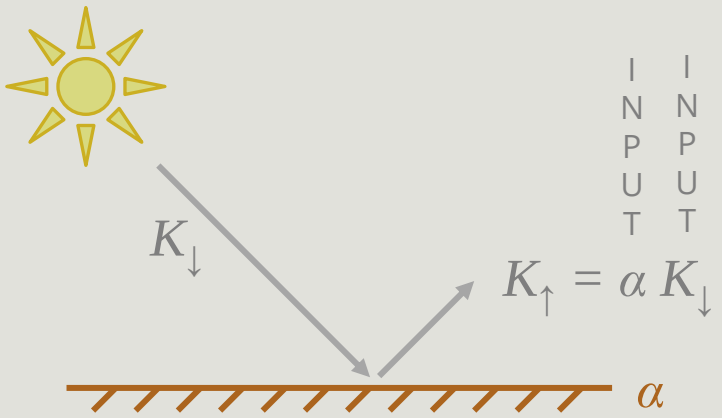
Salamanca et al. (2018) JGR



WHICH QUANTITIES TO EVALUATE?

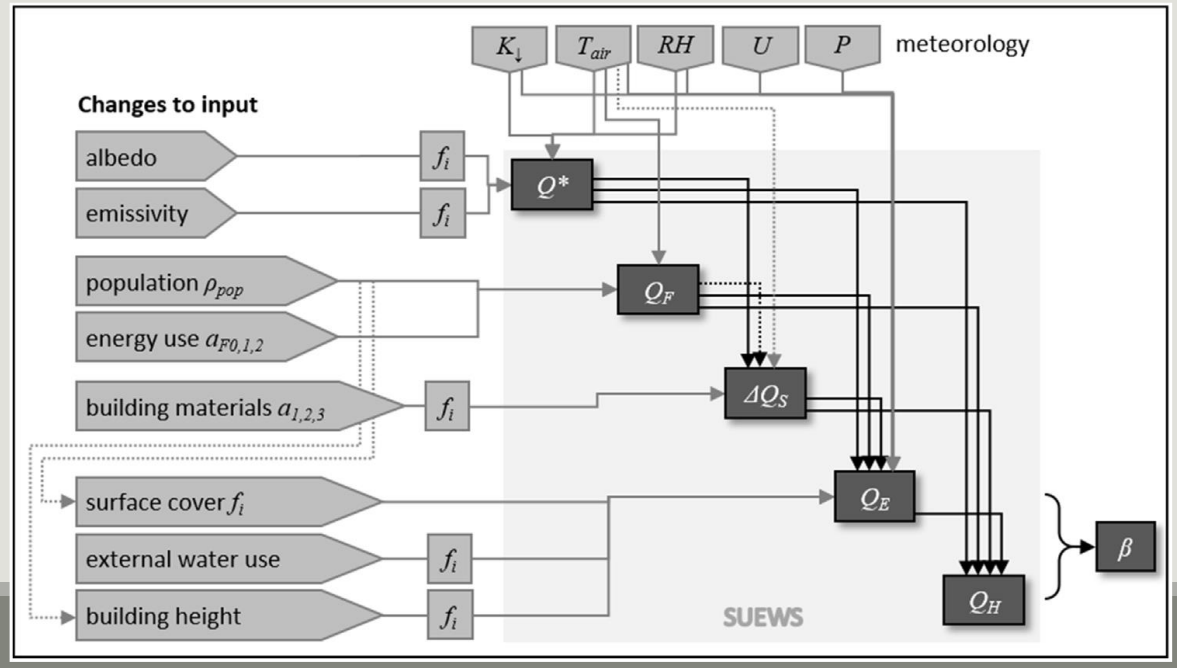
Theoretical perspective:

- What do we want to know about, and which processes/variables are relevant?
- What other processes/variables are relevant or could be relevant?
- What does the model need/what does it do to simulate these processes/produce this output?
- Which other variables can tell us about how the model is performing?



A flow diagram of model steps (at least for the variables evaluated) can be very useful to identify relevant variables and understand interdependencies

Ward & Grimmond (2017) LUP



WHICH QUANTITIES TO EVALUATE?

Practical perspective:

- Which quantities are modelled?
- Which quantities are available from observations?

} How compatible are these?

Modelled variables ⟷ Observed variables

Always consider the compatibility and independence of these pairs of variables

In some cases model output/observations can be simply adjusted to allow for a more direct evaluation…
… but in other cases modelled/observed variables may be very difficult to observe/model

- conversion between units
- simple conversions (e.g. E to Q_E, RH to q)
- more complex conversions (e.g. T_{sfc} to $L_{\uparrow}$ requires emissivity)
- derivation of T_{2m} from above-canopy temperature
- estimation of PBL height from observations
- very difficult to measure advection, but relatively straightforward to calculate from model grid

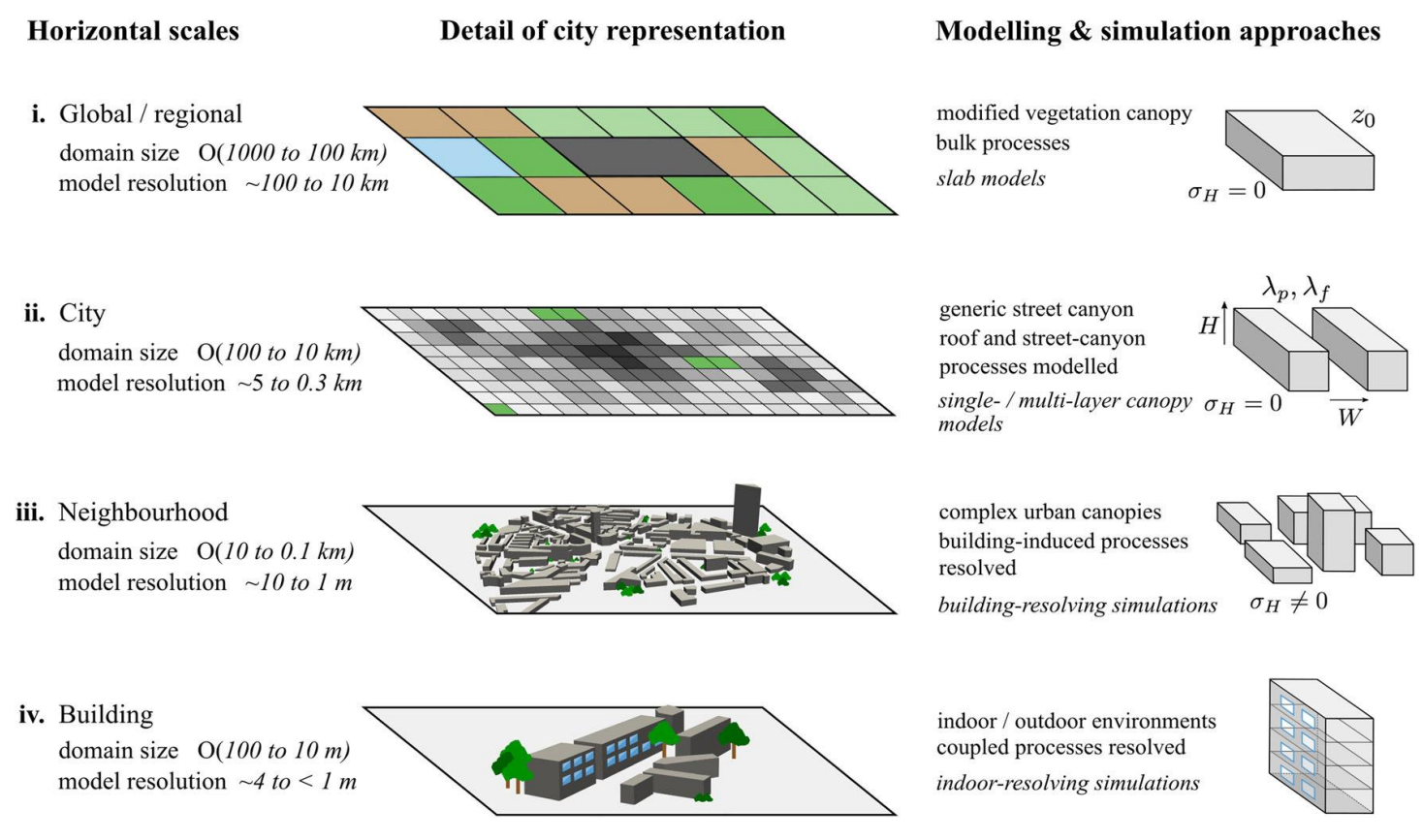
trivial — more involved → highly challenging

WHICH PROCESSES AND SCALES ARE MODELLED?

- How is the urban area represented?
- How is the urban canopy layer represented?
- Which processes/scales are parameterised and which are resolved?
- What processes/features are accounted for by the model?

WHICH PROCESSES AND SCALES ARE MODELLED?

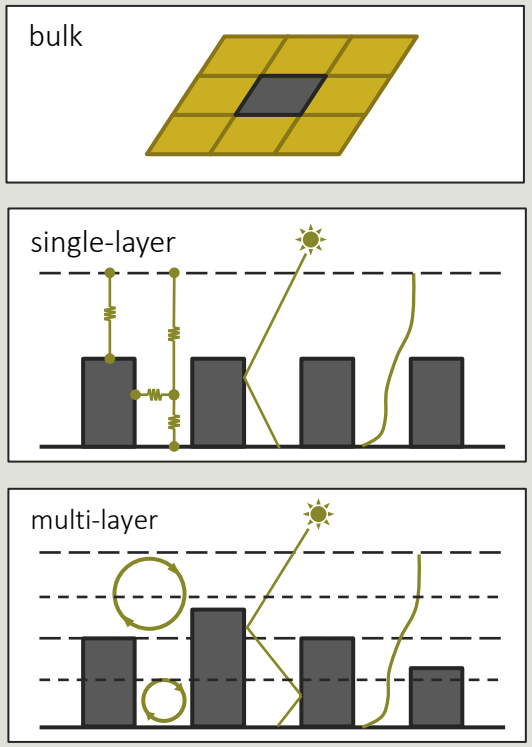
How is the urban area represented? Cities vs. neighbourhoods vs. buildings

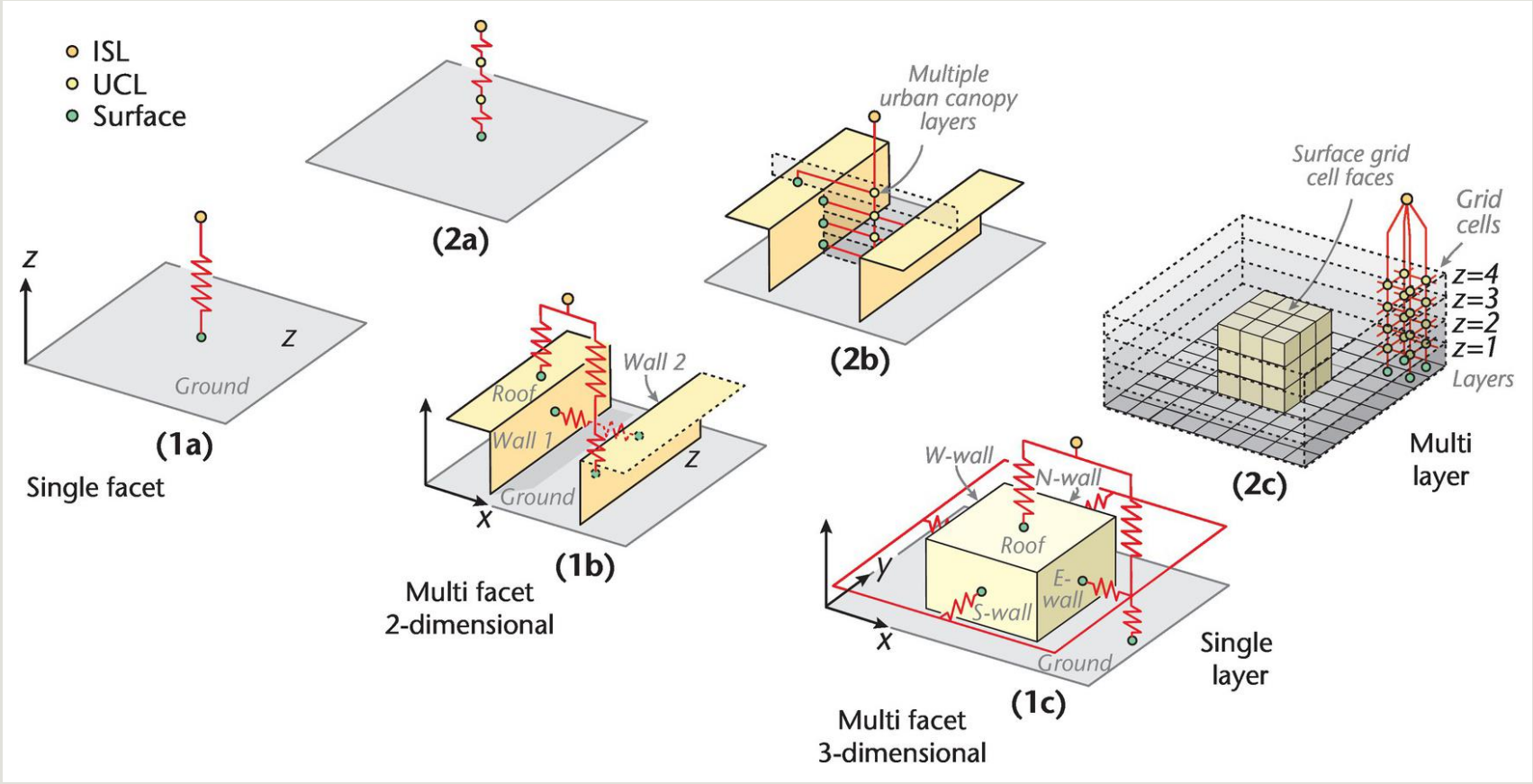


Hertwig et al. (2020) TAC

WHICH PROCESSES AND SCALES ARE MODELLED?

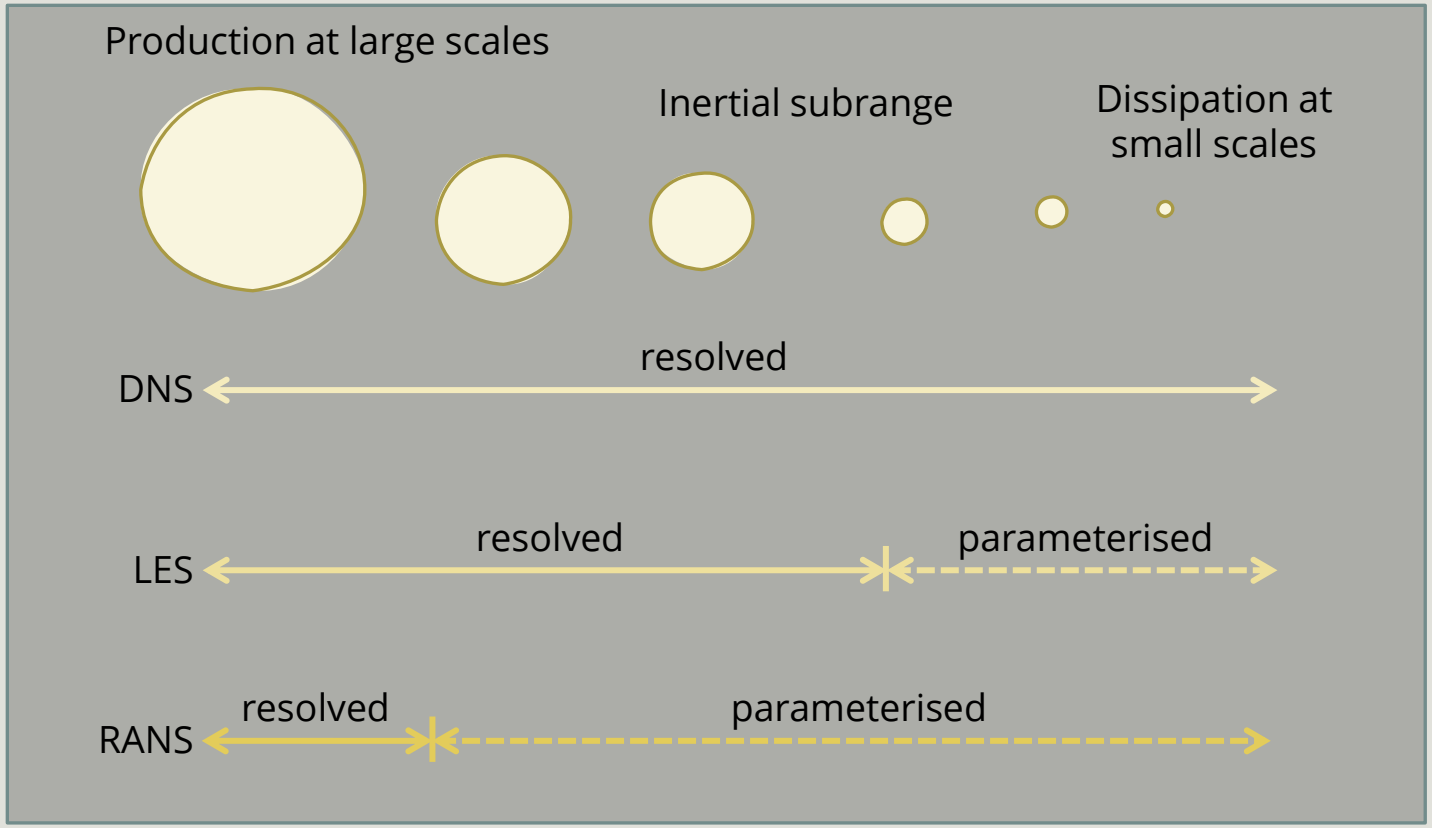
How is the urban canopy layer represented? Bulk vs. single-layer vs. multi-layer; surface vs facets vs resolved





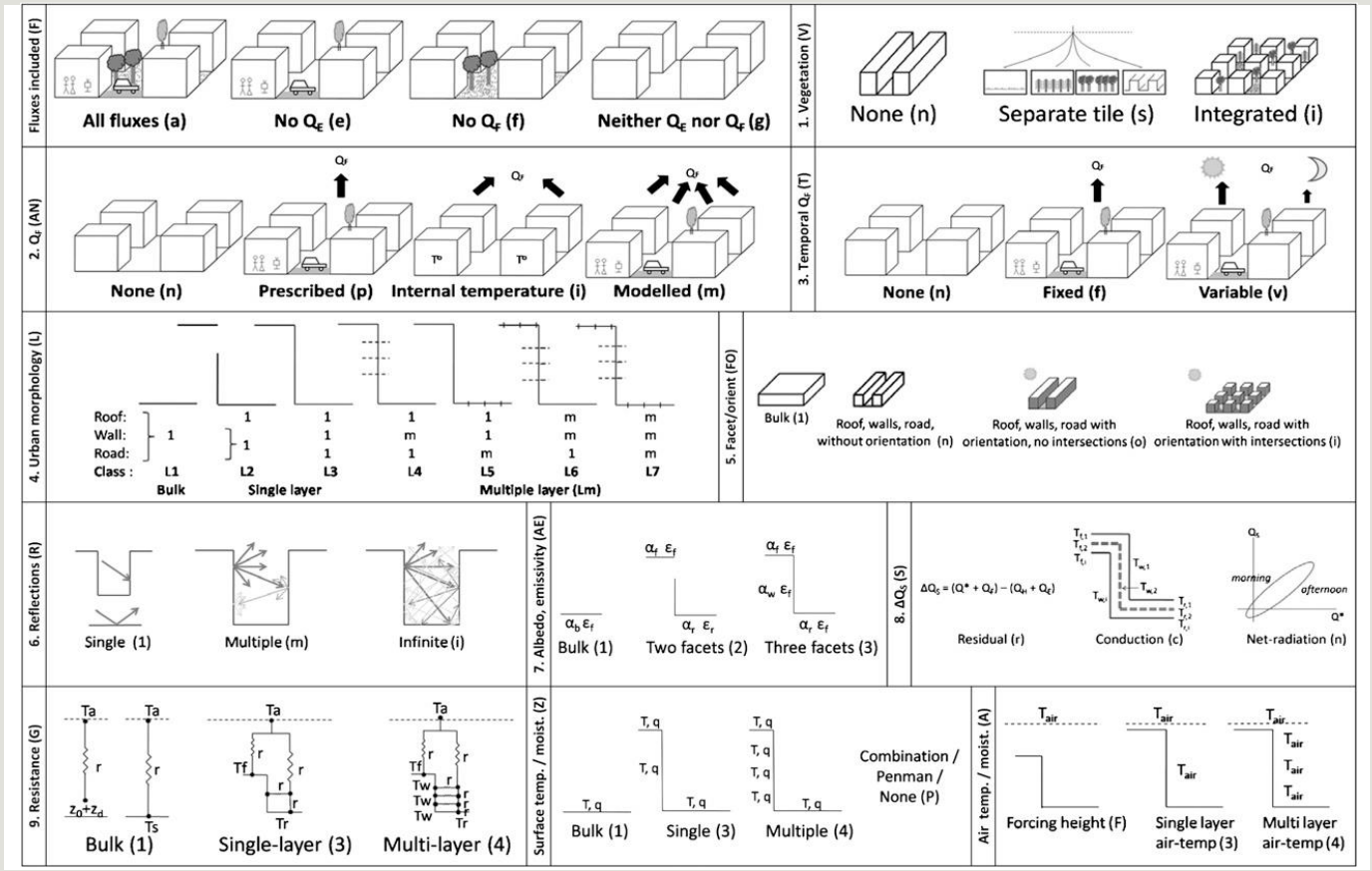
WHICH PROCESSES AND SCALES ARE MODELLED?

Which processes/scales are parameterised and which are resolved? i.e. size of grid box



WHICH PROCESSES AND SCALES ARE MODELLED?

What processes/features are accounted for by the model? Urban vegetation, anthropogenic heat, indoor climate

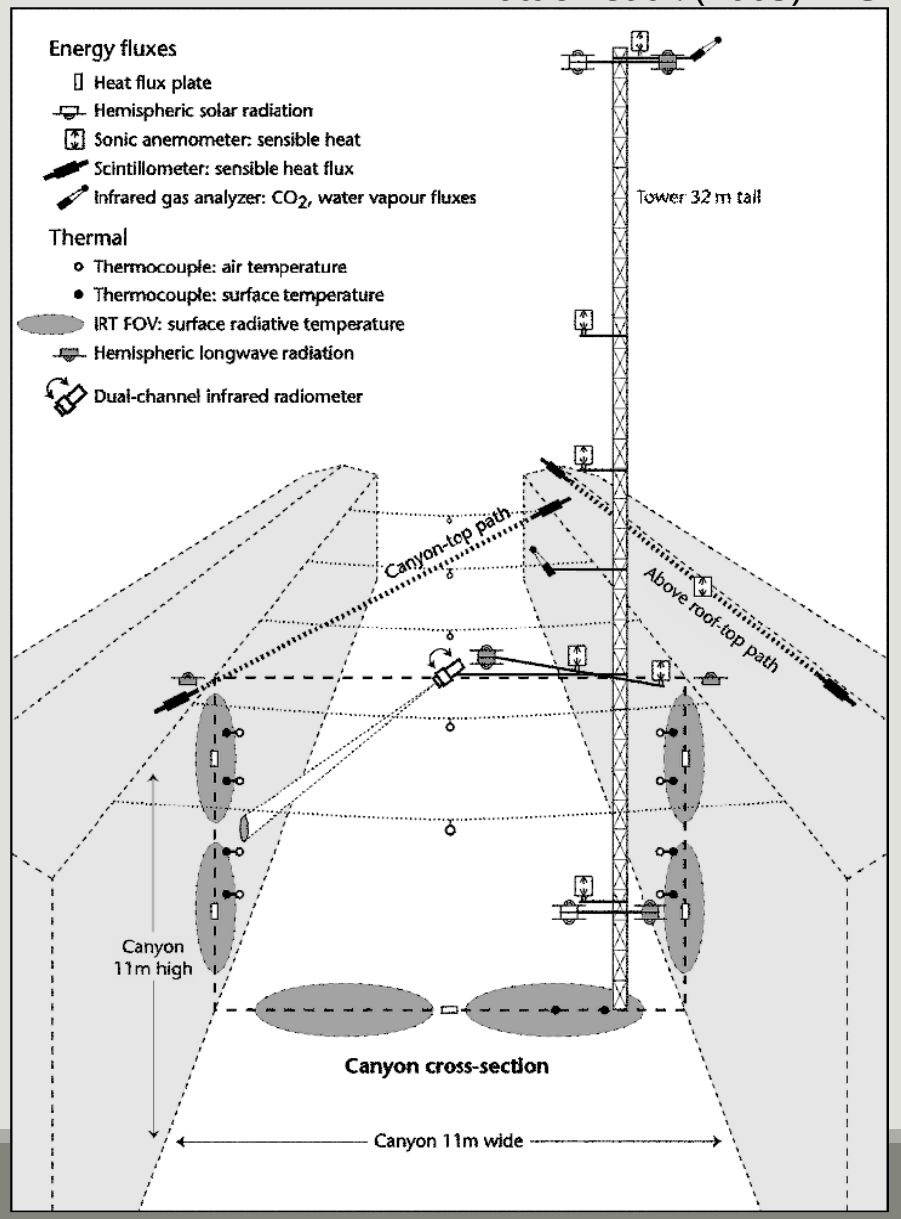


Grimmond et al. (2010) JAMC Characteristics used to classify urban energy balance models

ASSUMPTIONS AND UNCERTAINTIES IN THE REFERENCE DATASET

- Observations (or other reference datasets) are usually not perfect!
- It is critical that observations are quality controlled
- How is the observed quantity estimated? What assumptions are involved? Are there known biases?
  - e.g. does the observation technique require a homogeneous atmosphere?
  - e.g. are certain processes assumed negligible?
  - e.g. does the retrieval require other input data?
- What does the measurement tell us exactly?
  - which quantity is retrieved?
    - e.g. brightness temperature vs. surface temperature?
    - e.g. net exchange vs. emissions?
  - what does the measurement represent spatially?
  - what does the measurement represent temporally?
- What is the measurement uncertainty?

Rotach et al. (2005) TAC



SELECTION OF EVALUATION PERIOD AND AVERAGING INTERVAL

- The application and availability of the reference dataset likely determines the type of conditions tested:
  - Typical conditions, extreme event, heat stress, air pollution episode
  - Seasonal variability
  - One or more case studies

- Possible restrictions:
  - Finite reference dataset (e.g. duration of field campaign) and availability of data (after QC)
  - Limited computing resources (possible trade-off between period simulated and spatial resolution)
  - Availability and suitability of input data (e.g. met forcing data, surface characteristics)

- Are the sampling times/averaging intervals compatible between model and observations? Or is some averaging/recalculation necessary?

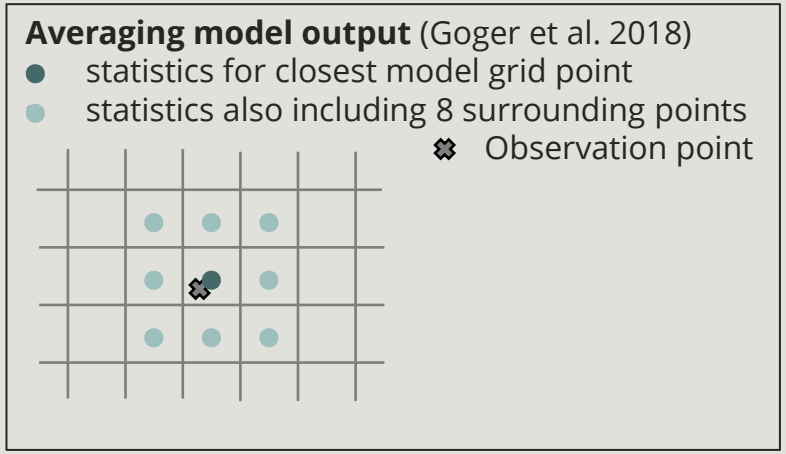
- It may be necessary to select remove modelled data where the reference data are missing, particularly for statistics

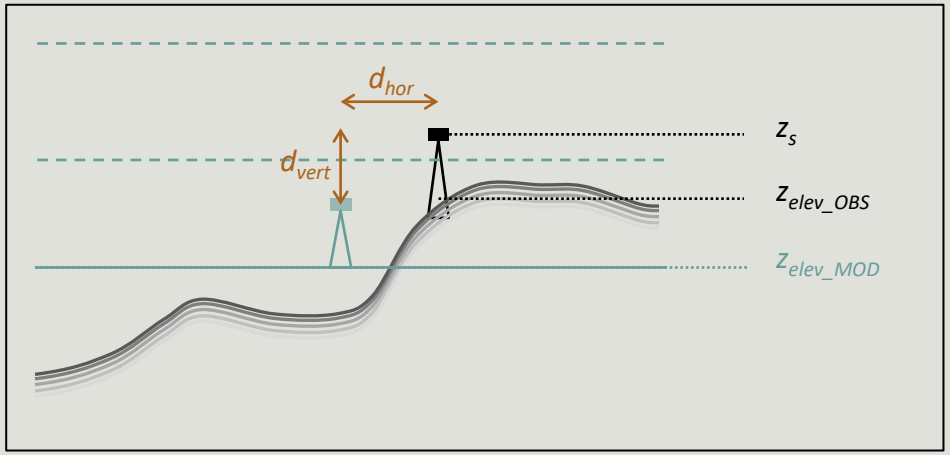
- Usually do not want gap-filled data (unless looking at totals) since gap-filling is usually done with another ‘model’ so is not a direct comparison to observations

SPATIAL SELECTION OF MODELLED (AND OBSERVED) QUANTITIES

- Need to consider how the spatial representativeness of the observations corresponds to model output
  - Vertically – which height do the observations represent? (Height above surface + altitude of site)
  - Horizontally – point, area-averaged or spatially distributed measurements?

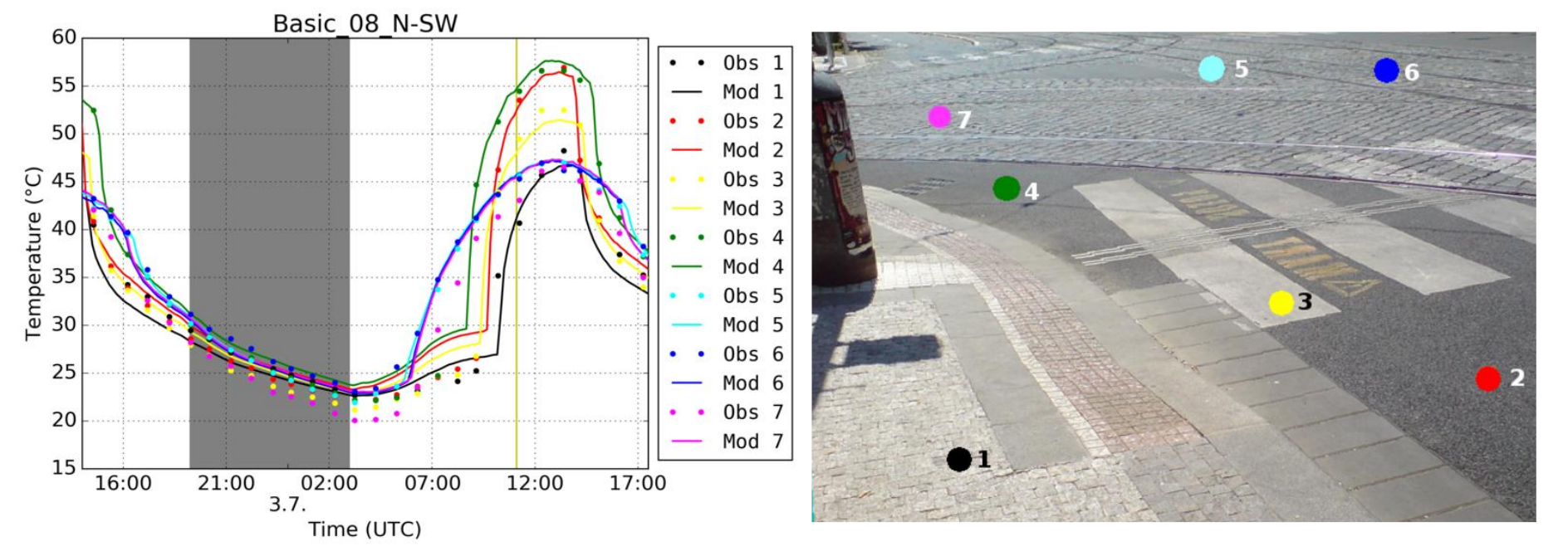
- How to best extract model data to compare with observations?
  - Vertically: closest height level above surface vs. interpolation
  - Horizontally: nearest grid point vs. nearest grid points vs. points with similar characteristics
  - Virtual measurements (footprint from model/obs, virtual scintillometer, virtual lidar)
    - allow a more consistent comparison of model and observations
    - enable investigation of measurement theory and insight into measurement uncertainties





EVALUATION OF GROUND TEMPERATURE FOR DIFFERENT SURFACE PROPERTIES

Resler et al. (2017) GMD: **surface temperatures during summer heat-wave event in Prague**

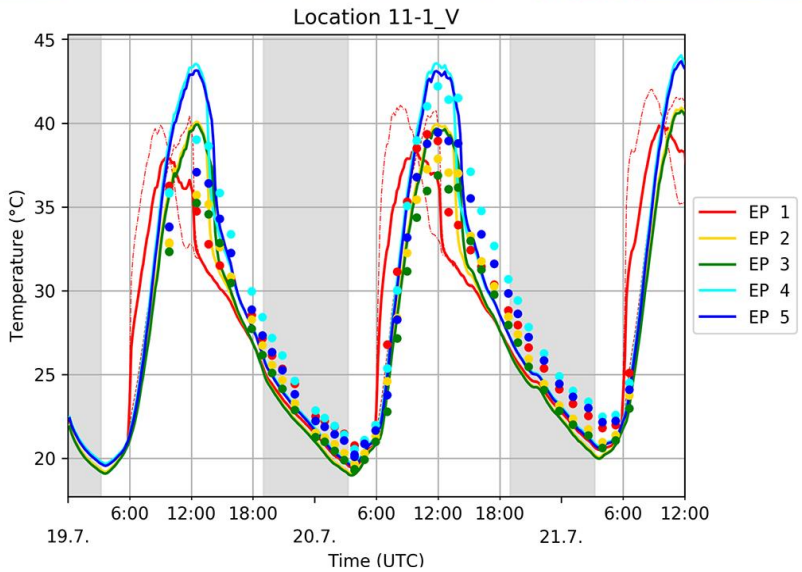


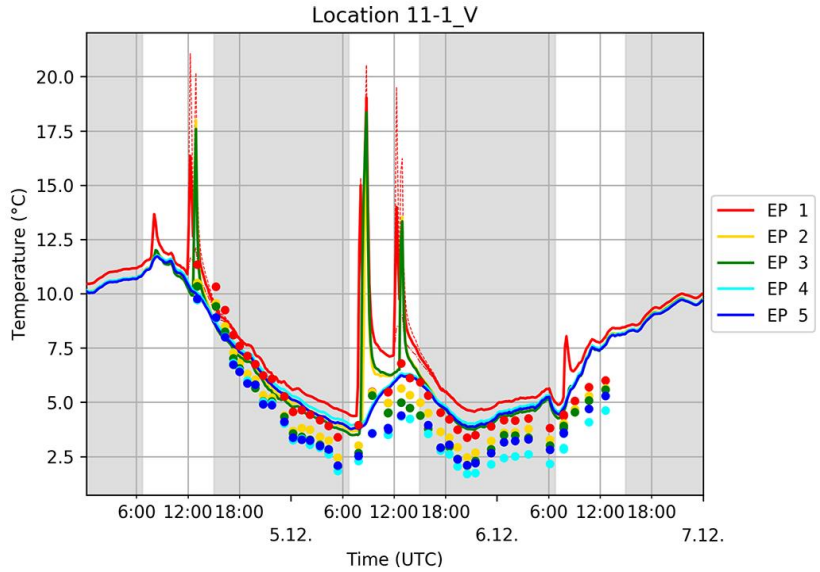
- Dark asphalt (points 2,4) have the highest maximum temperatures
- White painted asphalt (point 3) has a lower maximum temperature due to the higher albedo
- Paving blocks (points 1, 5,6,7) have lower maximum temperature and different diurnal cycle

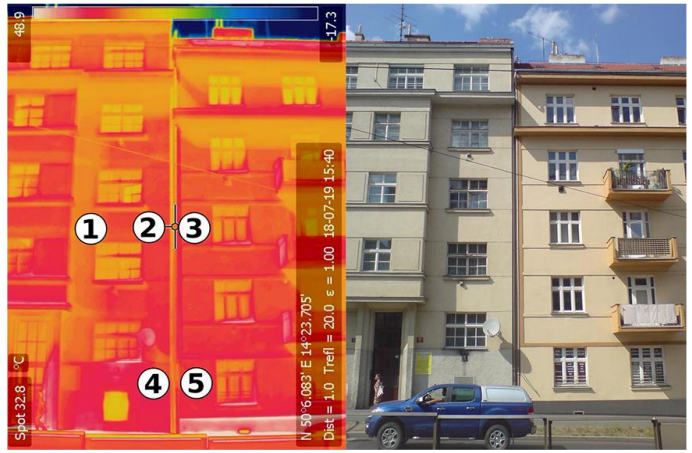
Ground temperatures
- derived from infrared camera (points)
- simulated by PALM at 2.08-m grid spacing (lines)

EVALUATION OF WALL TEMPERATURE AT DIFFERENT POSITIONS

Resler et al. (2021) GMD: **surface temperatures during summer and winter case-studies in Prague**







Wall temperatures

- hourly data from infrared camera (points)
- modelled data from PALM at 2-m grid spacing (lines)

- Good performance in summer; model overestimates observations during winter
- Sharp peaks in winter due to sun/shade pattern from buildings opposite
- Differences between observations and model attributed to differences in the timing of observations and model and imperfect model representation of fine-scale building structure

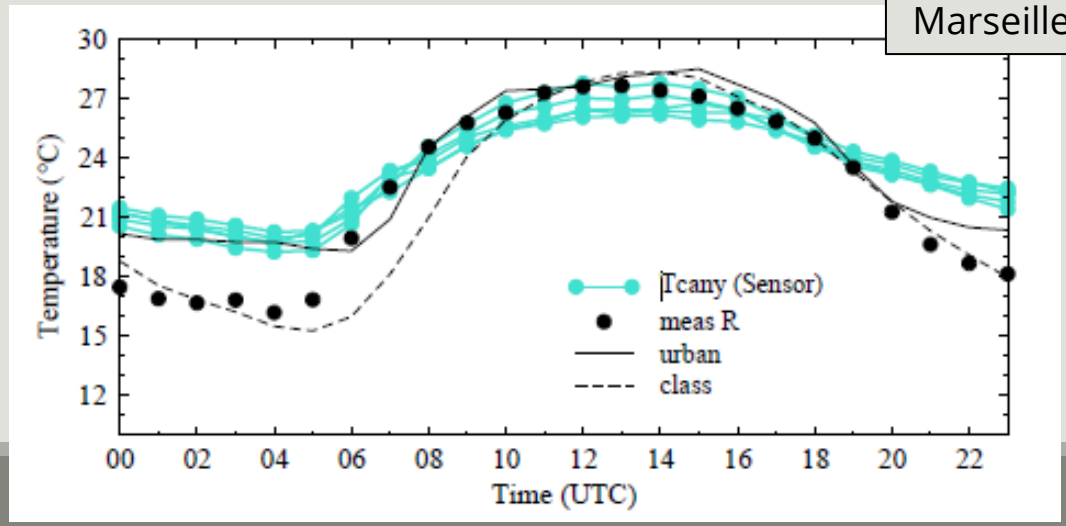
EVALUATION OF FACET TEMPERATURES FOR DIFFERENT ORIENTATIONS

Hamdi & Schayes (2007) ACP:

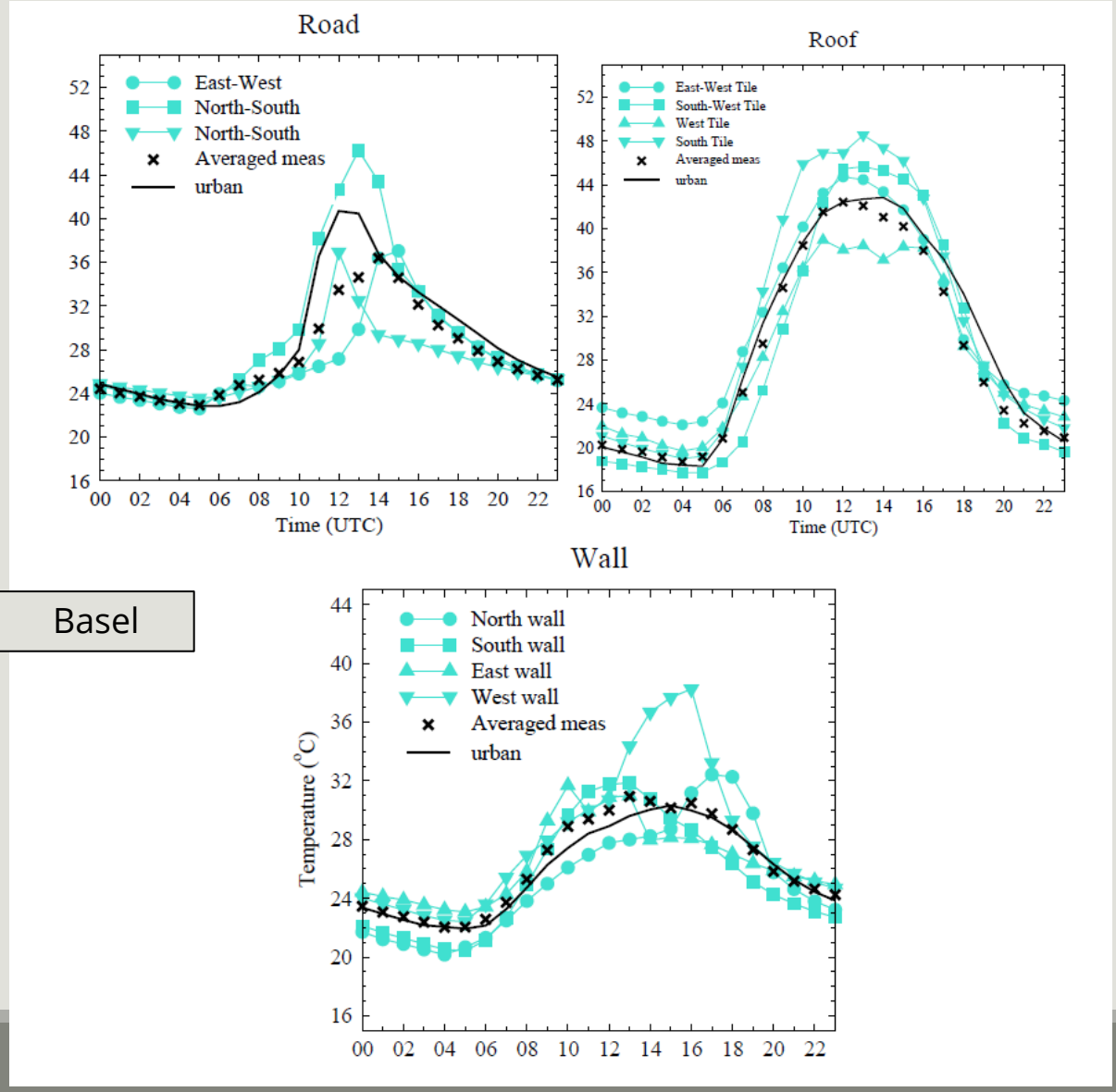
- **average road, roof and wall temperatures in summer in Basel**
- **average canyon temperatures in summer in Marseille**

- several microscale observations selected to be as representative as possible and then averaged
- comparison with WRF-BEP run on a single column
- data from field campaigns in two different cities used

Marseille



Basel

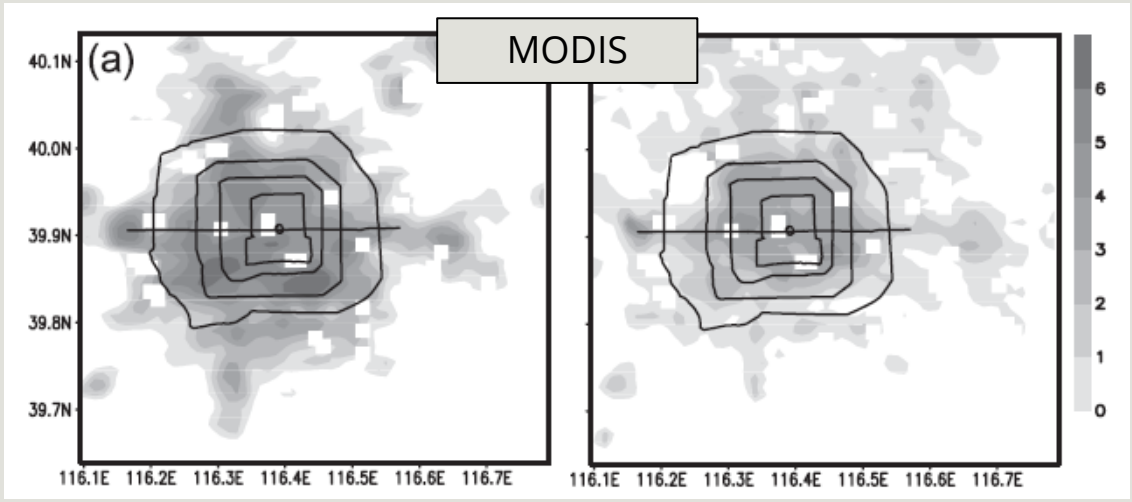


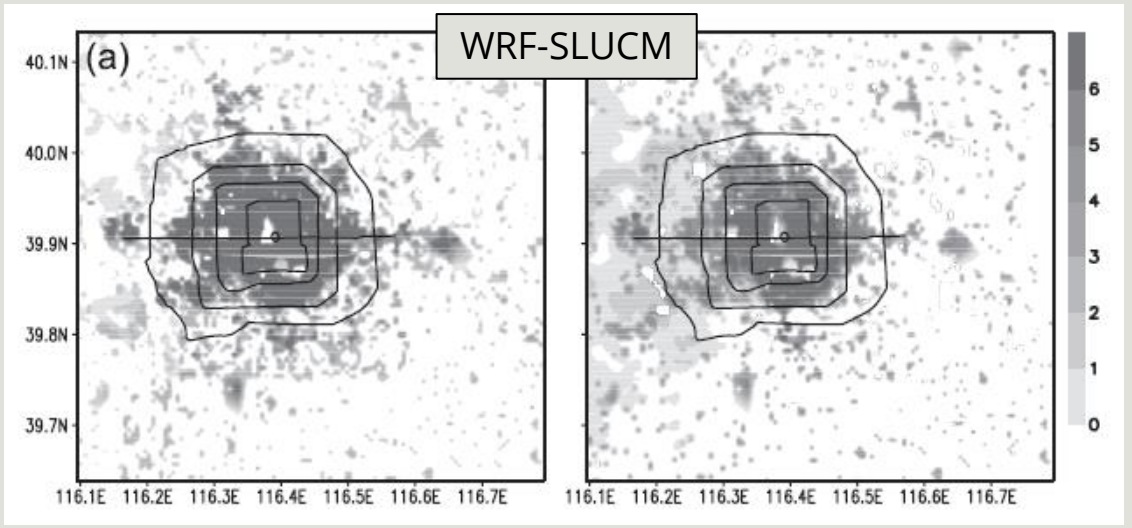
EVALUATION OF SURFACE-UHI USING LST

Miao et al. (2009) JAMC: **surface urban heat island in Beijing**

- 1-km MODIS LST
- WRF-SLUCM at 500-m resolution

- WRF captures the spatial extent of the SUHI and location of maximum SUHI reasonably well
- But WRF overestimates MODIS LST by about 3 °C at night





EXAMPLES

EVALUATION OF WIND AND TURBULENCE

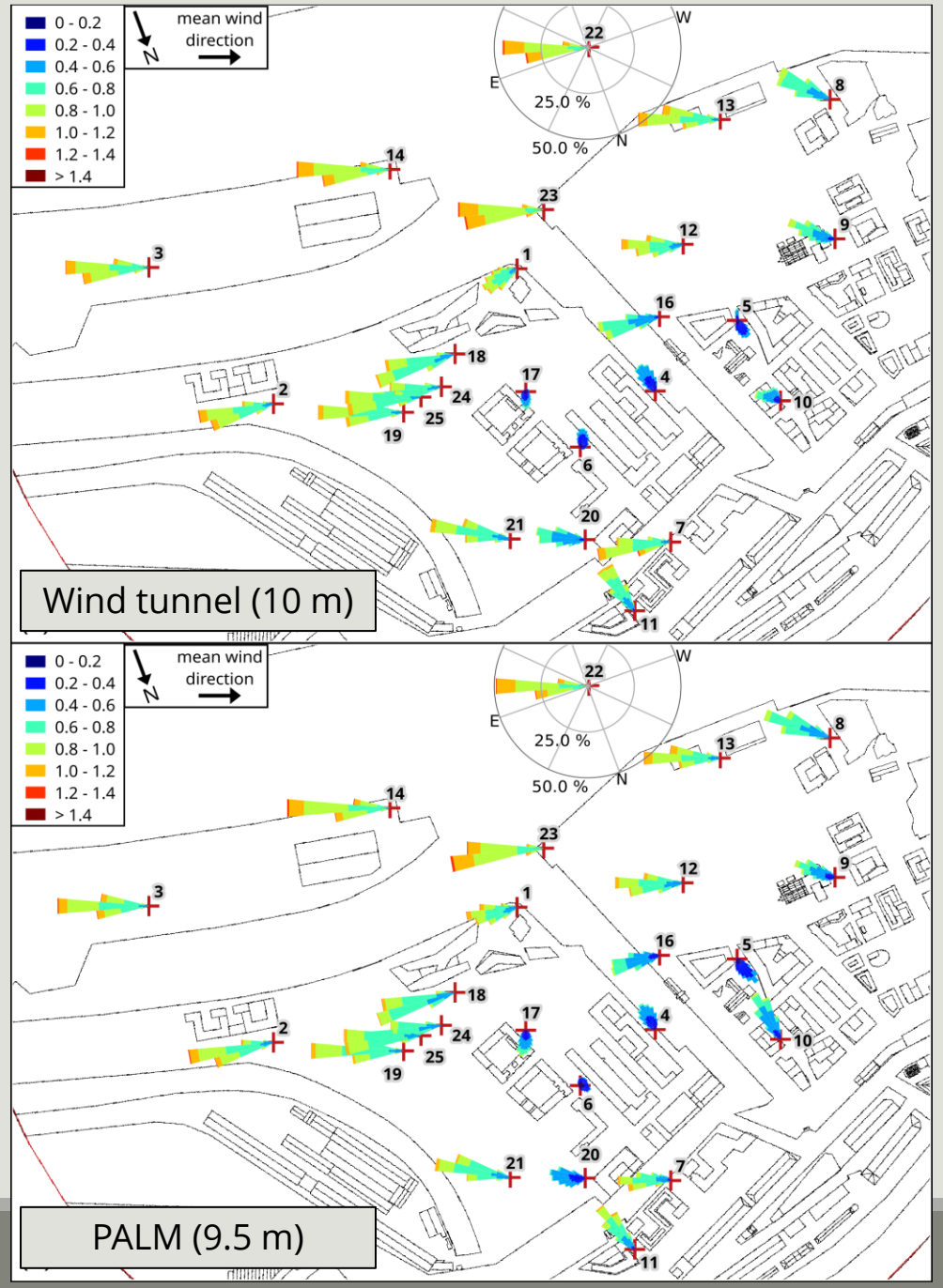
Gronemeier et al. (2017) GMD:

- **Wind and turbulence measured at 25 locations in wind tunnel with 1/500 scale model of HafenCity, Hamburg**
- **Evaluation against PALM dynamical core with 1-m grid spacing**

- Wind tunnel allows evaluation of airflow without temperature or humidity effects

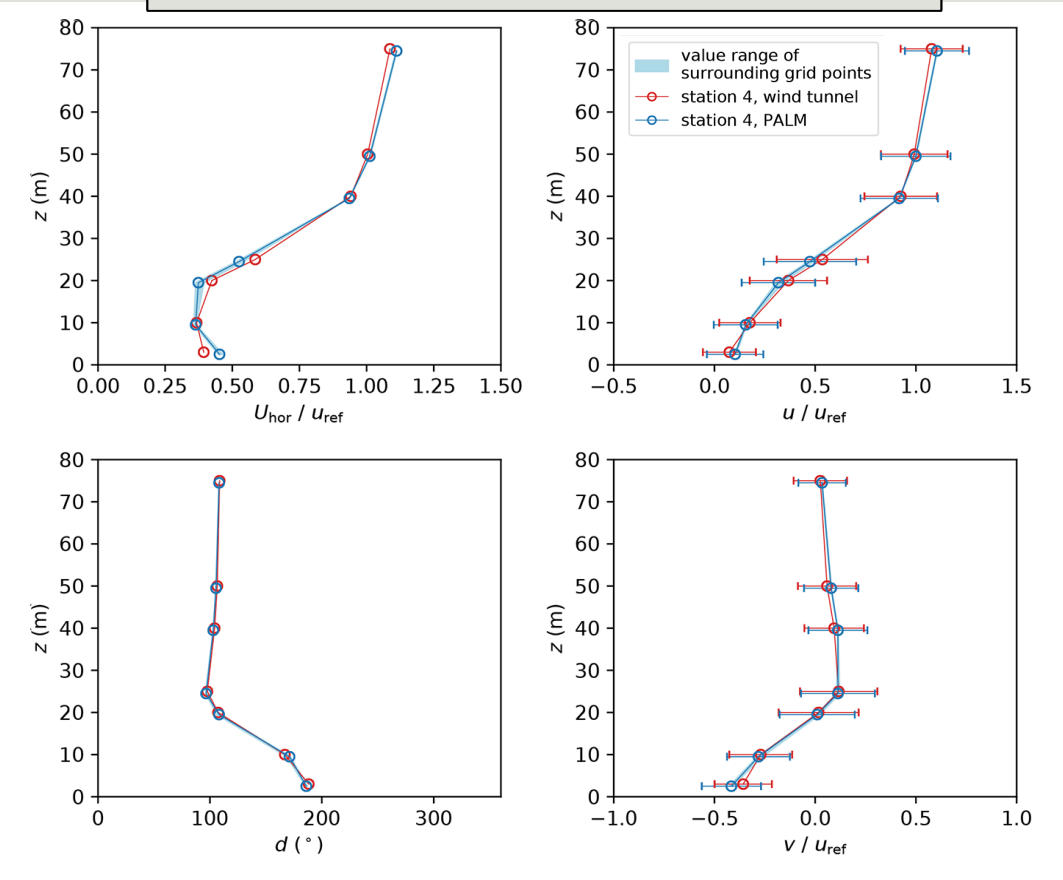
- Difficulties matching wind tunnel locations to model grid points
  - horizontally (PALM grid points at slightly different locations (< 1m) and with slightly different topography; neighbouring grid points also considered)
  - vertically (PALM levels 0.5 m below wind tunnel levels; no adjustment done so as not to introduce additional uncertainty)

- Qualitative and quantitative comparisons

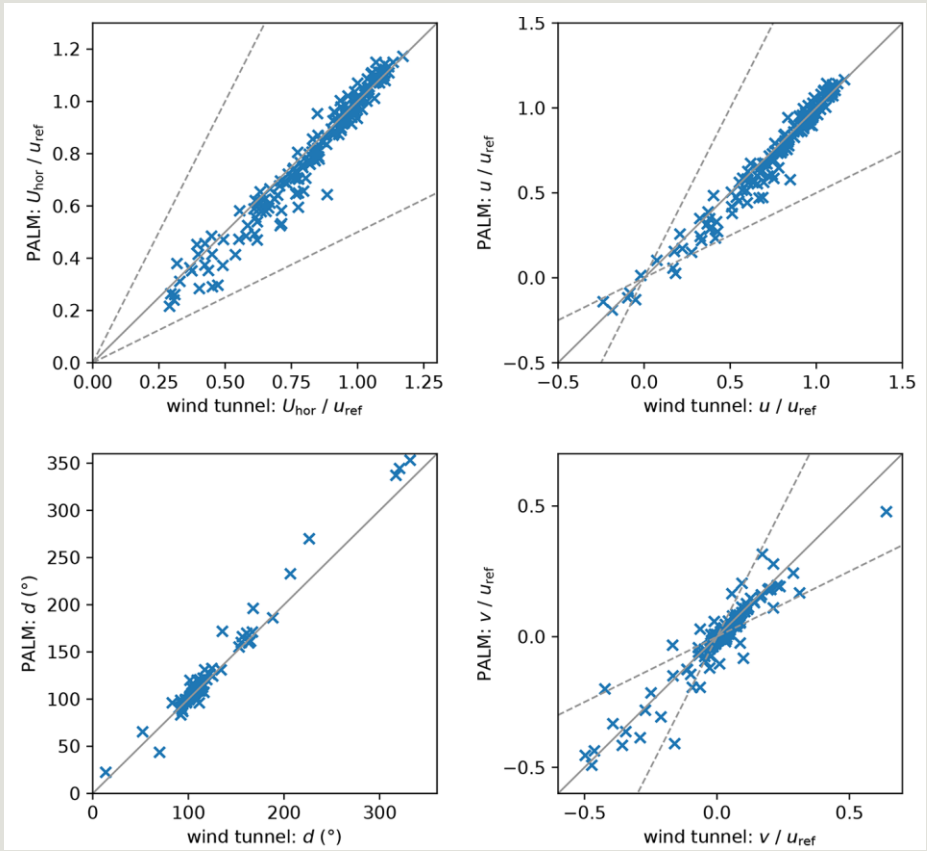


EVALUATION OF WIND AND TURBULENCE

Station 4 (less complex building structure)



Metric	u/u_{ref}	v/u_{ref}	U_{hor}/u_{ref}	σ_u^2/u_{ref}^2	σ_v^2/u_{ref}^2	I_u	I_v	Ideal
FAC2	0.98	0.73	1	0.98	0.98	1	1	1
H	0.91	0.70	0.96	0.82	0.79	0.93	0.91	1
R^2	0.97	0.87	0.96	0.57	0.55	0.83	0.85	1
FB	–	–	0.03	-0.06	0.19	-0.08	0.03	0
MG	–	–	1.05	0.95	1.2	0.93	1.04	1
NMSE	–	–	0.01	0.07	0.21	0.05	0.07	0
VG	–	–	1.01	1.05	1.08	1.02	1.02	1
δ_a	0.025	0.025	0.025	0	0	0	0	



Gronemeier et al. (2017) GMD

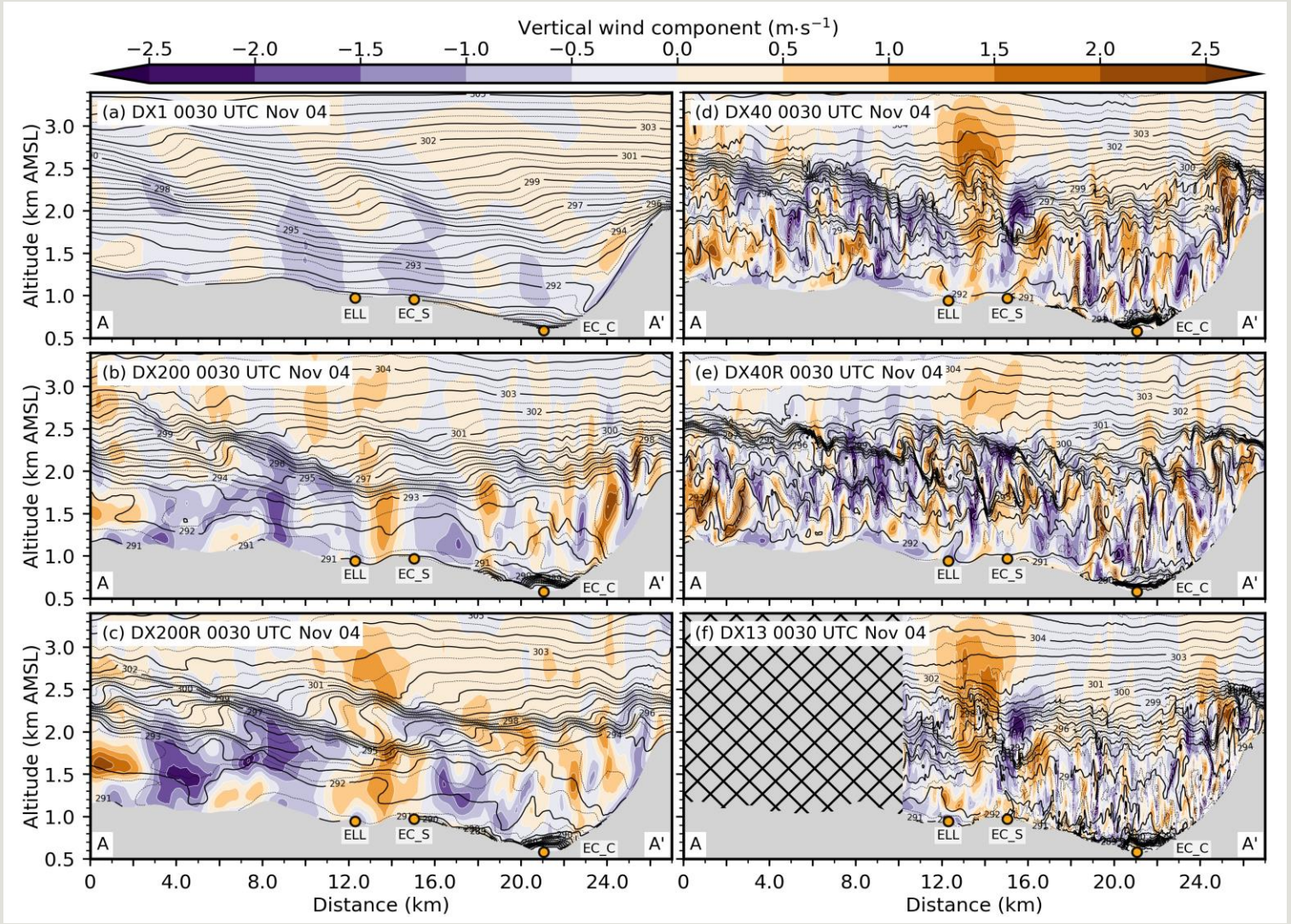
COMPARISON OF PROCESSES AT DIFFERENT GRID RESOLUTIONS

Umek et al. (2022) QJRMS:
Foehn case study over the city of Innsbruck

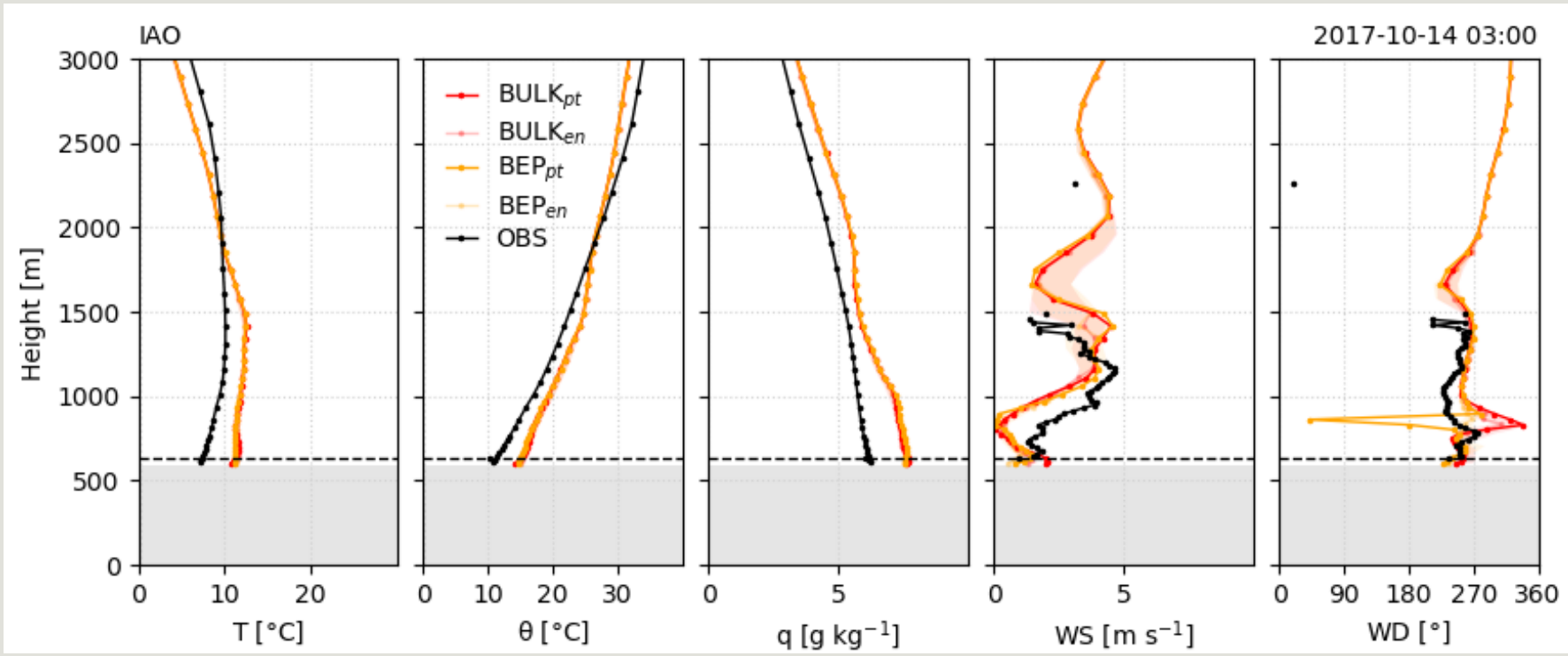
- WRF mesoscale (1 km) simulations
- WRF-LES (200 m, 40 m, 13 m) simulations

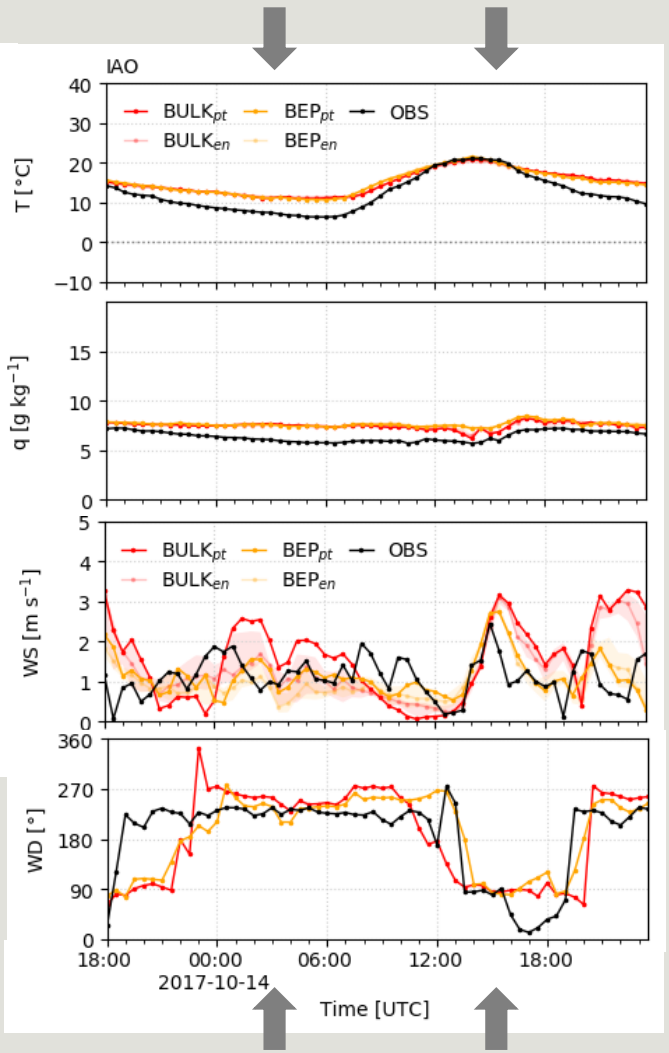
- Observations (ground-based, profiles) used to guide interpretation since highly complex situation not replicated by model

- Mesoscale simulation cannot capture fine-scale features
- LES does a much better job
- Finer scale LES improves the representation of turbulence but does not have a large impact on mean quantities



EVALUATION OF BOUNDARY LAYER PROFILES





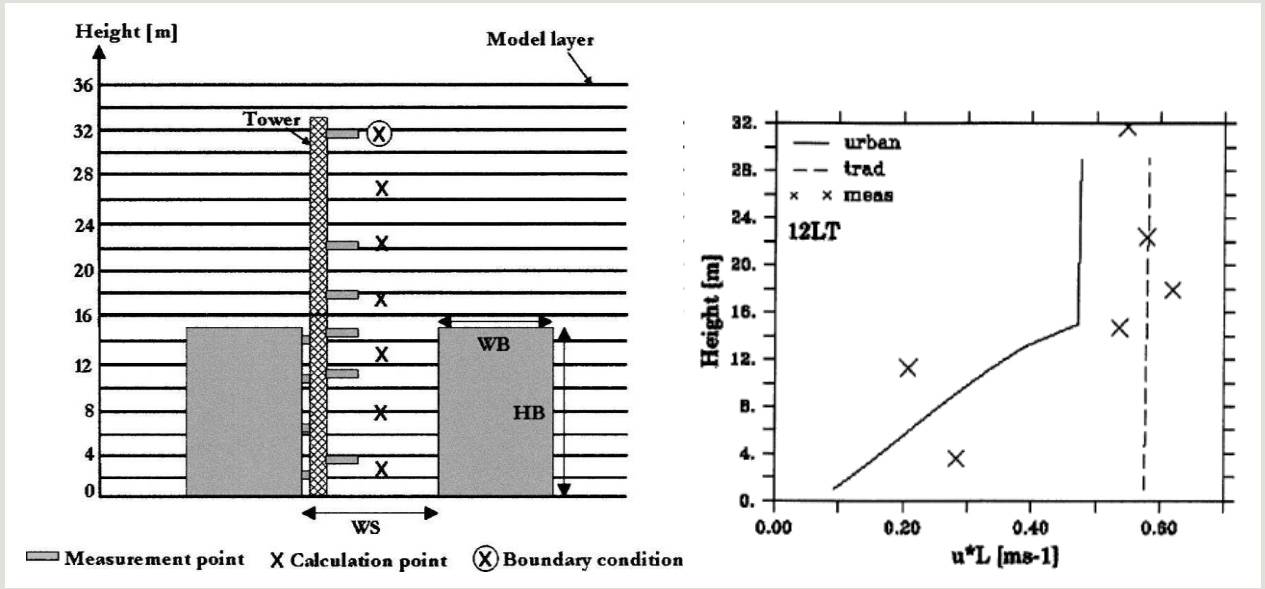
Valley wind **case over the city of Innsbruck**

- WRF mesoscale (1 km) simulations with bulk and BEP scheme
- Observed ABL profiles with microwave radiometer and Doppler wind lidar
- Near-surface temperature modelled well during the day but overestimated at night
- Humidity modelled reasonably well but overestimated at night
- Vertical structure of winds and diurnal cycles well captured

EVALUATION OF MULTI-LEVEL URBAN CANOPY MODEL AGAINST FLUXES

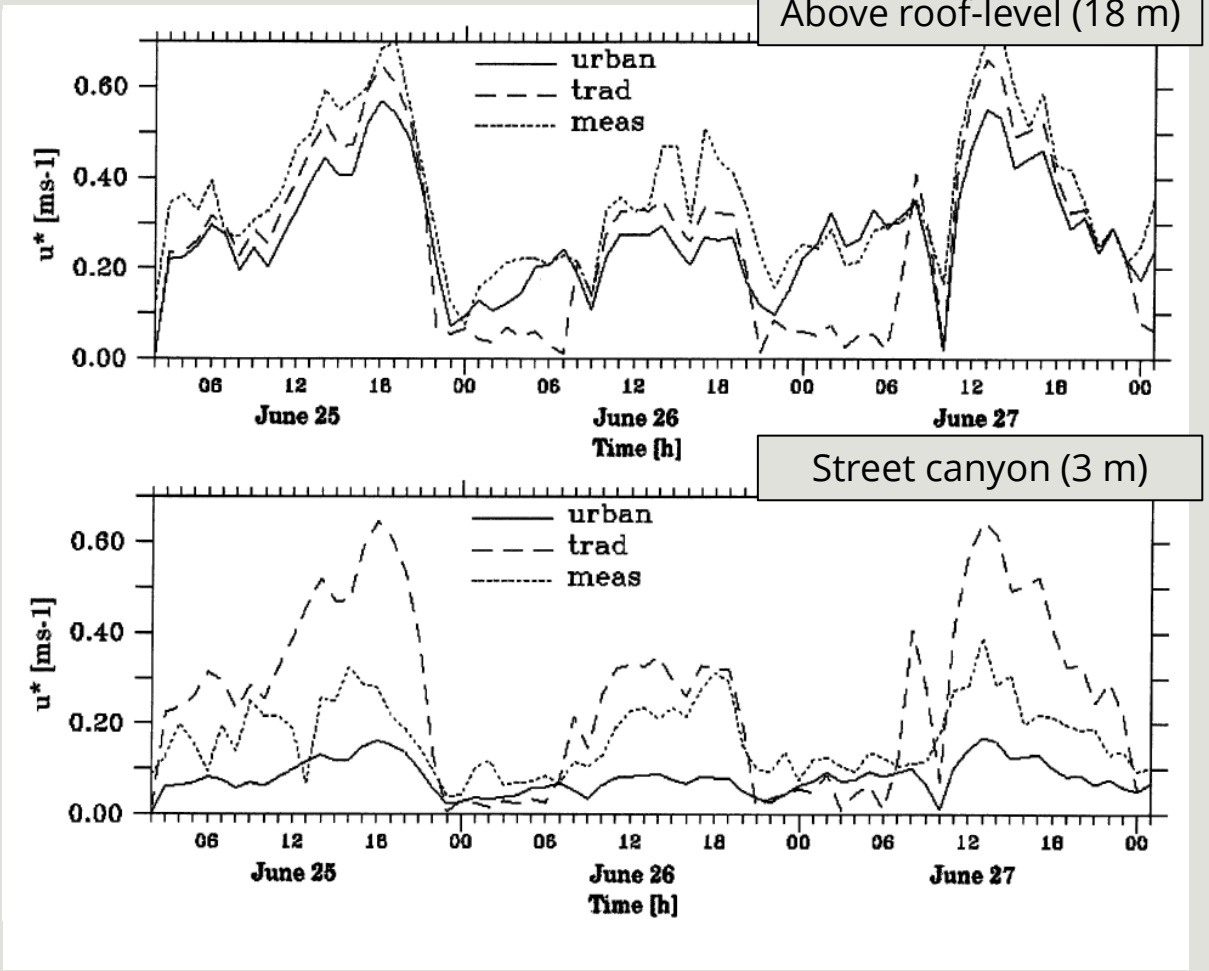
Roulet et al. (2005) JAM:

- **BUBBLE dataset for Basel provides multiple quantities at multiple levels within and above the UCL**
- **Comparison with WRF-BULK and WRF-BEP (single-column)**



Above roof-level (18 m)

Street canyon (3 m)

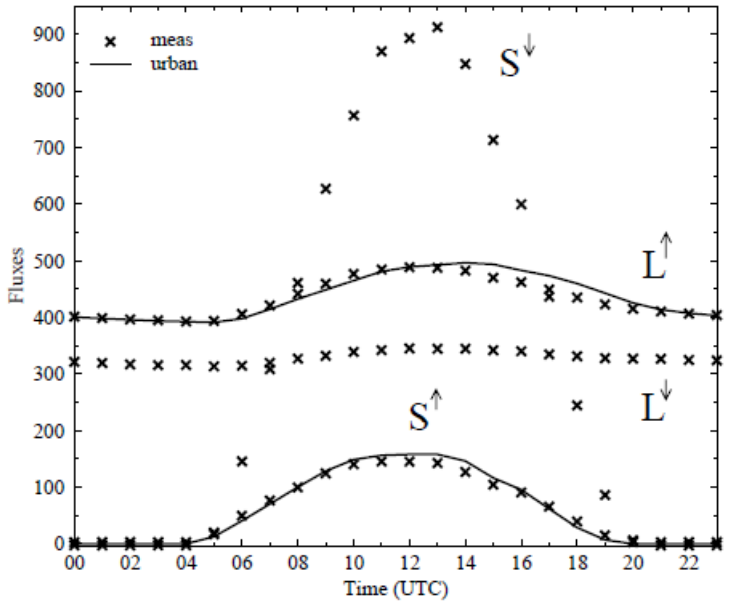


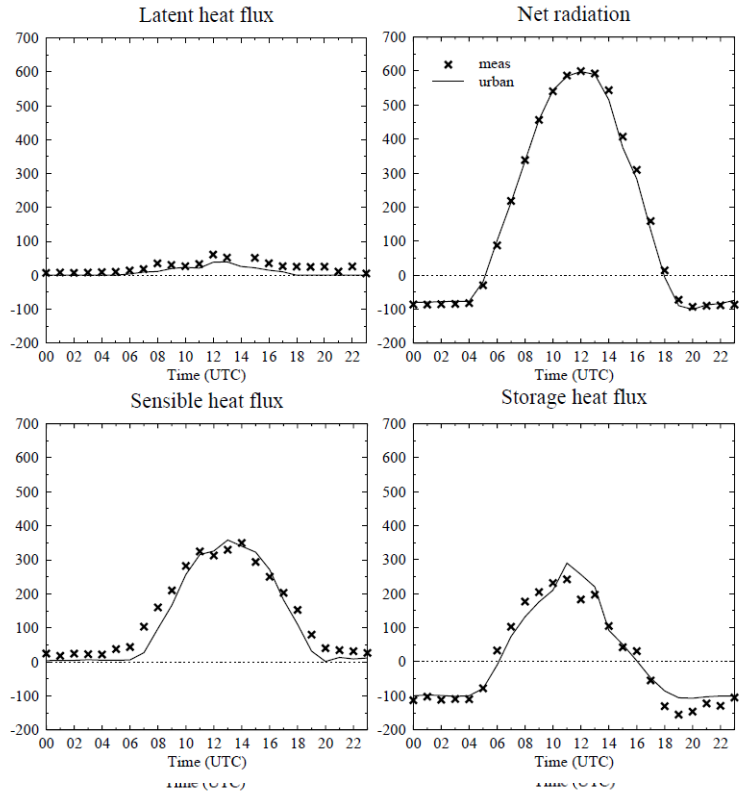
EVALUATION AGAINST RADIATIVE AND TURBULENT FLUXES

Hamdi & Schayes (2007) ACP:

- **Local-scale radiative and turbulent fluxes in summer in Basel and Marseille (top-of-tower measurements)**
- **comparison with WRF-BEP run on a single column**

Marseille





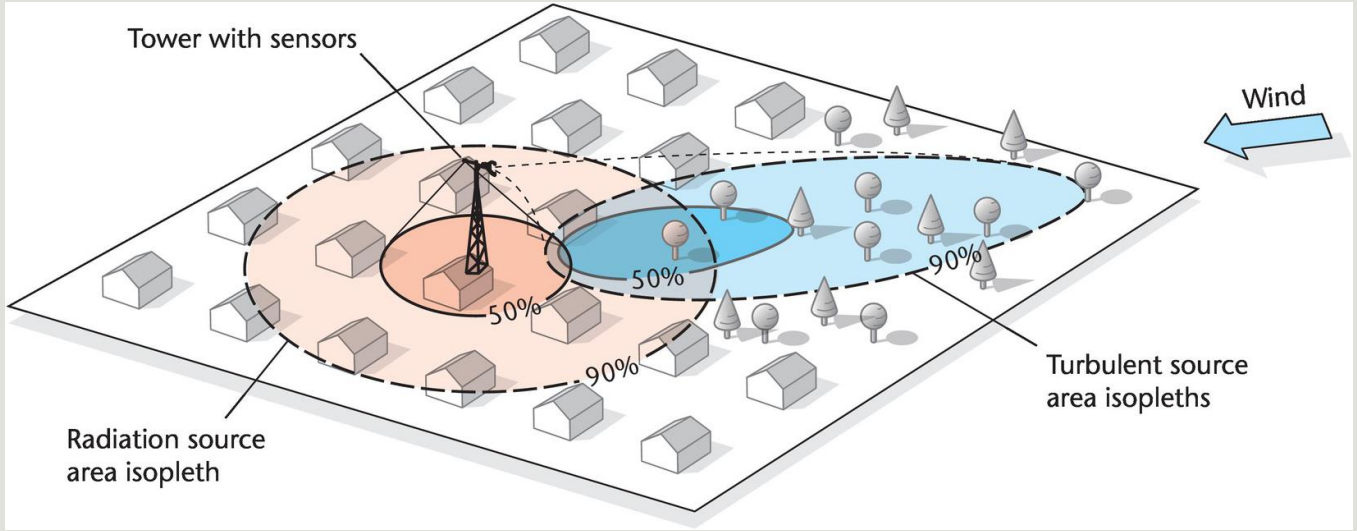
RADIATIVE AND TURBULENT FLUX FOOTPRINTS VS MODEL GRID

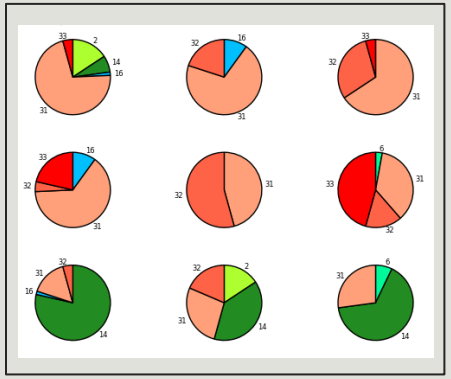
- Measurement footprint of radiometer and eddy covariance measurements differ
  - Radiometer footprint is circular and constant with time
  - Flux footprint varies with atmospheric conditions, particularly wind direction

- Homogeneous sites:
  - Footprint composition may be similar and similar to the composition of the grid box

- Heterogeneous sites:
  - Radiometer footprint composition may be different to the flux footprint composition
  - Flux footprint composition may vary in time
  - Both may have a different composition to the model grid box

Urban Climates





WRF land cover around Innsbruck site

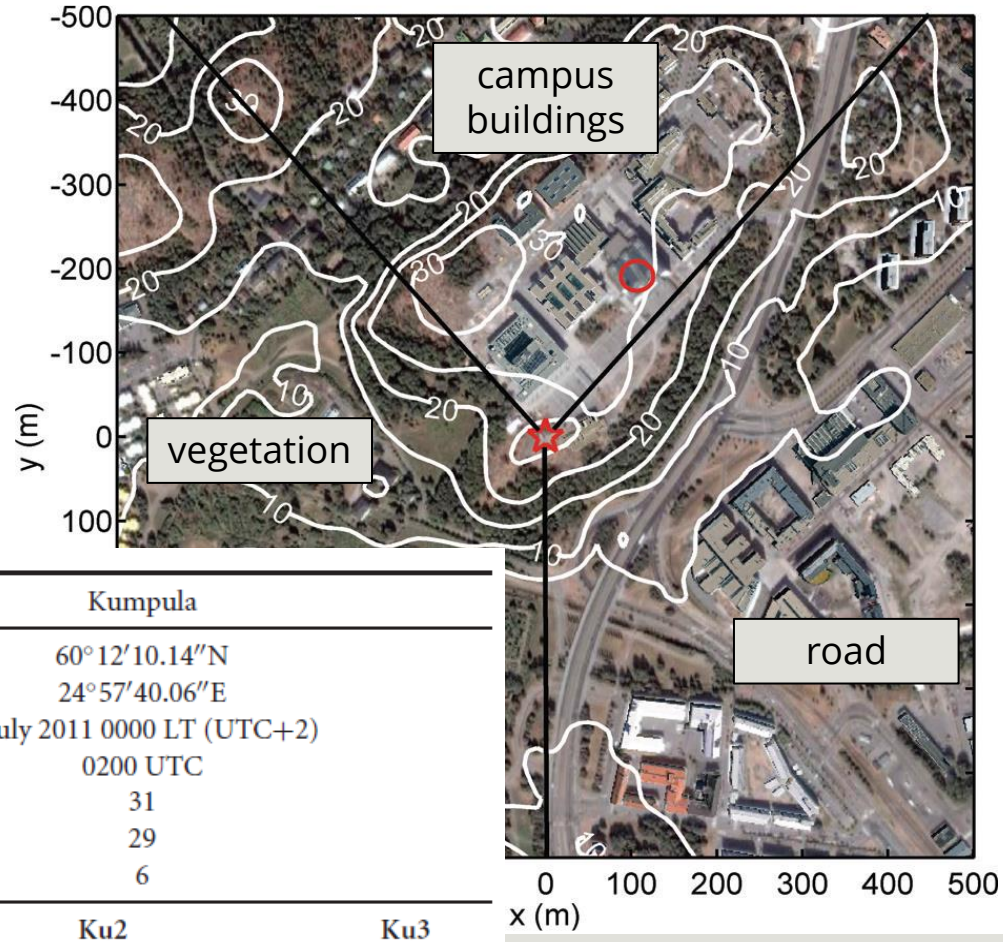
- Dryland cropland and pasture
- Cropland/woodland mosaic
- Evergreen needleleaf forest
- Mixed forest
- Water bodies
- Low intensity residential
- High intensity residential
- Industrial or commercial

EVALUATION AGAINST EDDY COVARIANCE AT A HETEROGENEOUS SITE

Nordbo et al. (2012) Tellus B

Karsisto et al. (2015) QJRMS:

- **Multi-seasonal evaluation of CLM, SURFEX and SUEWS models using observations at two sites in Helsinki**
- Kumpula site heterogeneity handled by modelling three distinct sectors separately and then combining the model output in a single timeseries based on wind direction (for turbulent flux data)
- For radiative flux data model output is considered from the most representative sector



	Torni	Kumpula
Latitude (WGS84)	60°10′04.09″N	60°12′10.14″N
Longitude (WGS84)	24°56′19.28″E	24°57′40.06″E
Initial date and time of the run	1 July 2011 0000 LT (UTC+2)	1 July 2011 0000 LT (UTC+2)
Time zone	0200 UTC	0200 UTC
Measurement/modelling height (m)	60	31
Base elevation (m)	15.2	29
Local climate zone (LCZ)	2	6

		Ku1	Ku2	Ku3
Study area (m^2)	1 960 000	447 000	782 000	782 000
Population density (no. m^{-2})[a]	0.0081	0.0031	0.0037	0.0044
Fraction of built surface	0.78	0.42	0.54	0.46
Fraction of paved surface	0.40	0.27	0.39	0.32
Fraction of buildings	0.37	0.15	0.15	0.14
Fraction of vegetation	0.22	0.58	0.46	0.54

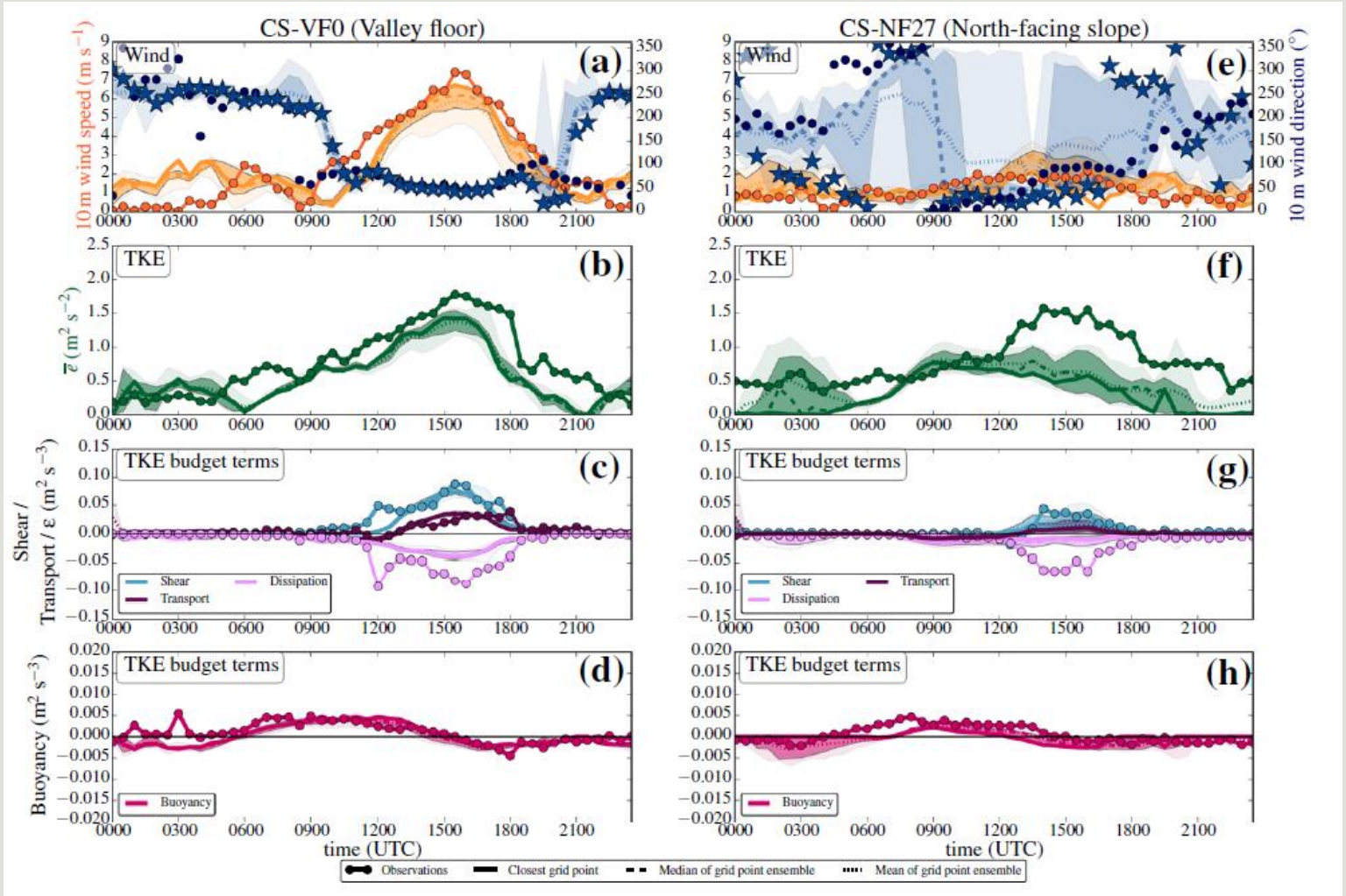
EVALUATION IN COMPLEX TERRAIN

Goger et al. (2018) BLM:

- **COSMO model evaluated against turbulent fluxes in the Inn Valley**
- Model grid point and nearest neighbours used to assess spatial variability
- Spatial variability (shading) higher for sloped site than for valley floor site

Umek et al. (2021) QJRMS:

- Cold bias for mountain-top potential temperatures partly attributed to height difference between model orography and the observations (approx. 300m)

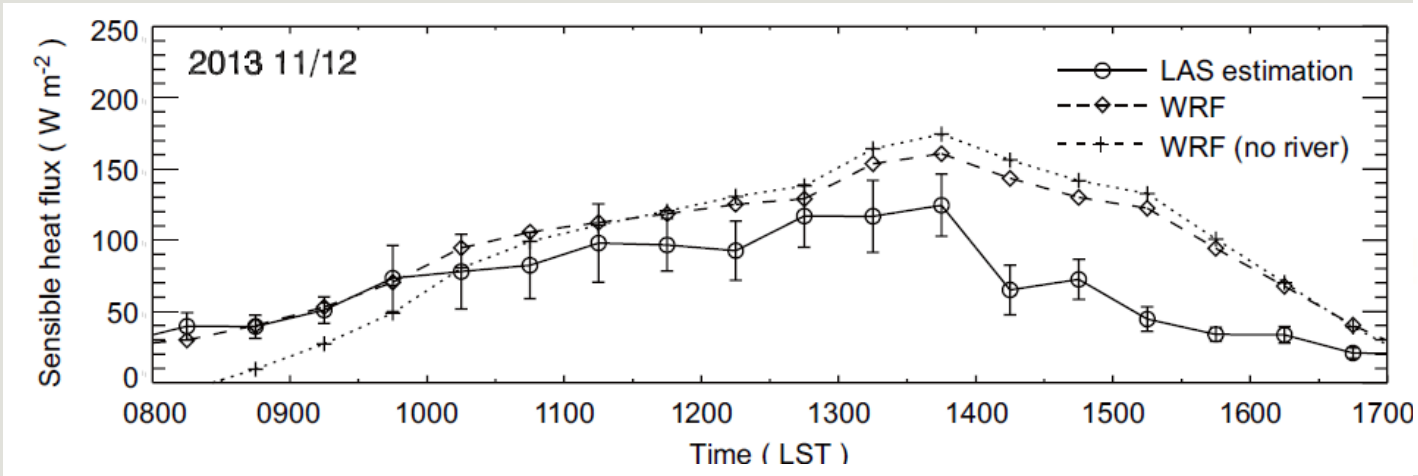


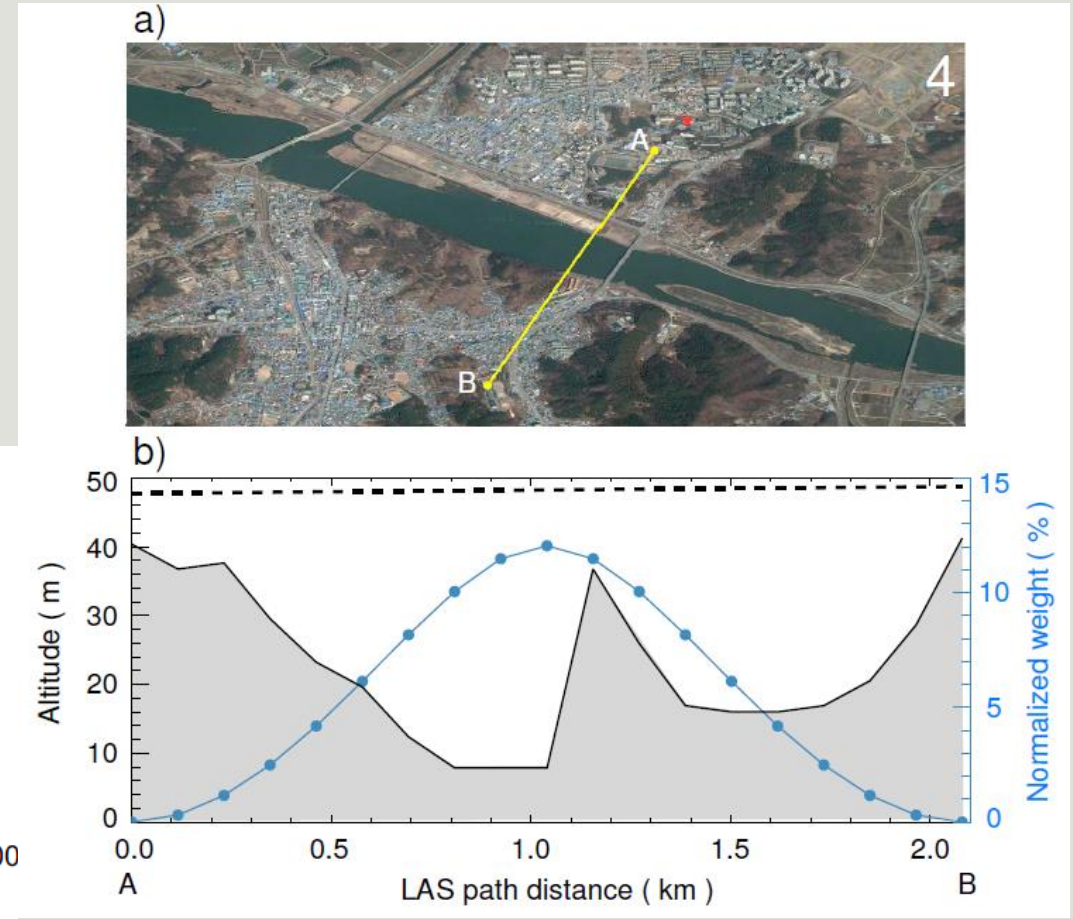
EVALUATION OF PATH-WEIGHTED MODEL OUTPUT AGAINST SCINTILLOMETRY

- Scintillometers provide measurements over large areas, often comparable to the size of a mesoscale model grid box

Lee et al. (2015) AAS:

- **Long scintillometer path in South Korea spanning several model grid boxes (WRF 200-m)**
- Averaging over multiple grid boxes accounts for weighting function of the instrument





IMPORTANT CHECKS AND THINGS TO AVOID

- Good understanding of both the model and the reference dataset is crucial (unless strong collaboration provides the necessary expertise on both sides)
  - Read references relating to the reference dataset
    - Need to understand interpretation and representativeness. Are there any oddities?
  - Read references relating to model
    - Need to understand how output is produced and model capabilities. Are there expected biases?
- Do multiple comparisons - plot timeseries, scatter plots before jumping to totals/statistics
- Check input conditions
- Check surrounding landscape and characteristics in the model, ideally against a reference dataset if available
- Avoid spin-up period and locations close to domain boundaries
- Sanity checks: check model converges, check resolved and SGS proportions, check order of magnitudes

STATISTICS – DANGERS OF LINEAR REGRESSION!

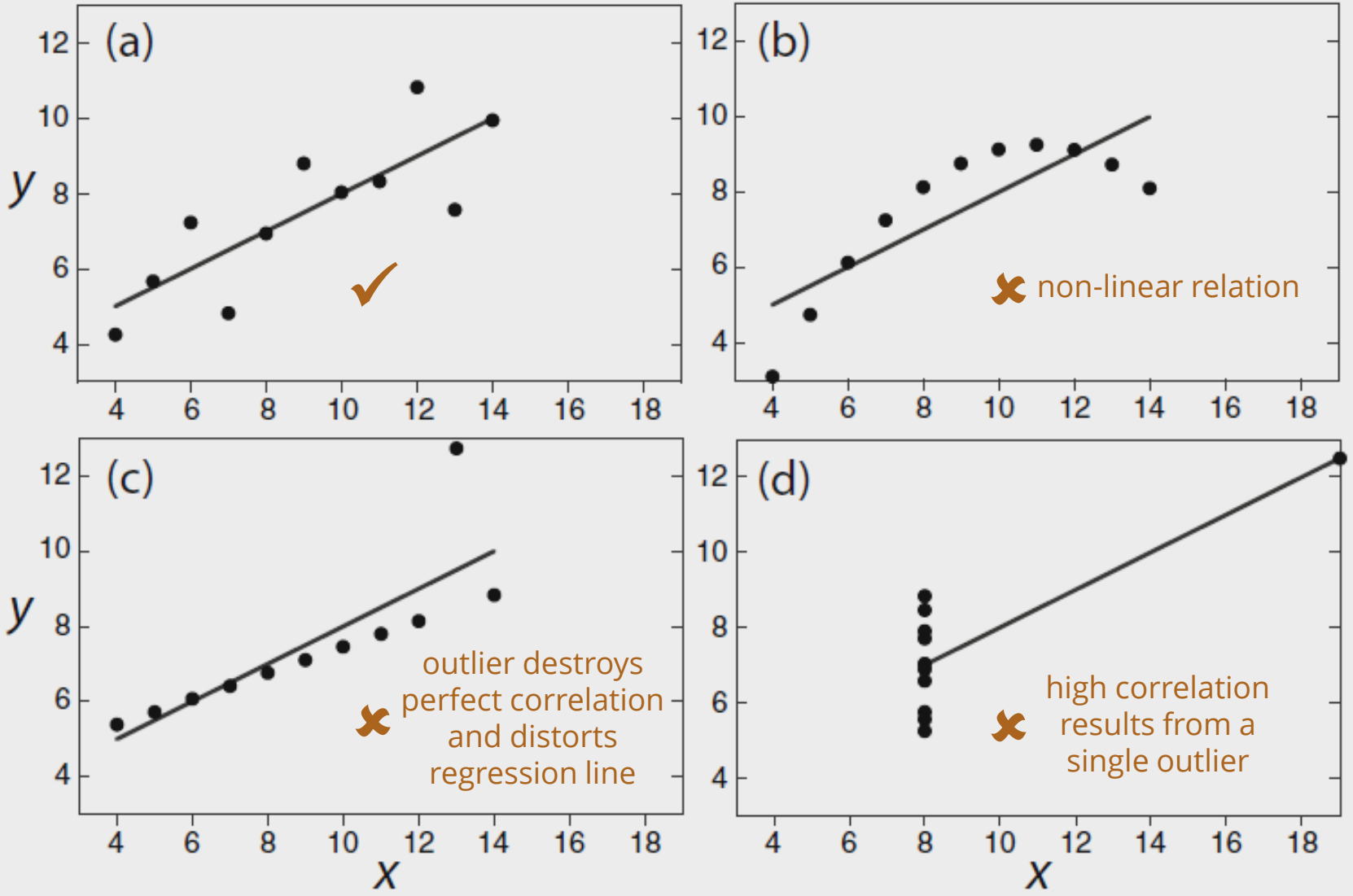


Fig. 11.12 Anscombe's (1973) quartet: four examples of datasets that have the same mean (7.5), same standard deviation (4.1), same linear regression line ($y = 3x + 0.5$), and same correlation coefficient (0.82).

Caution

- ! Only fit linear regression if makes sense
- ! With standard linear regression errors in y-variable minimised (but not x)

GOOD PRACTICE

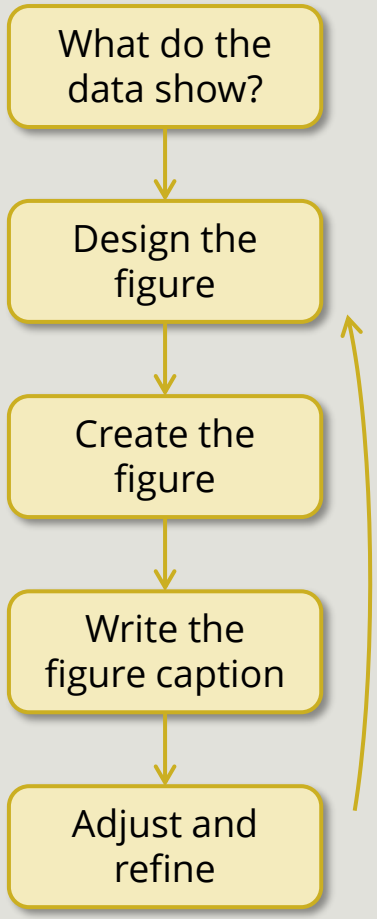
FIGURES

- Designing figures
  - what do the data show?
  - what is the clearest way to communicate this to the reader?
  - which aspects should be highlighted, which are less important?

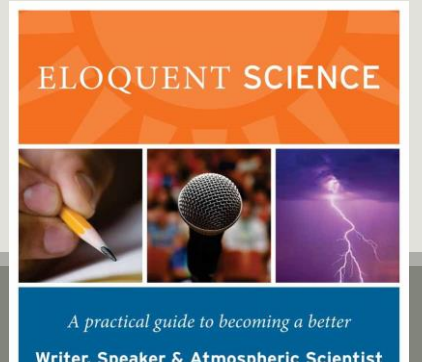
- Colours can have a huge impact on interpretation
  - If applicable, try to stick to convention (e.g. red – sensible heat flux, blue – latent heat flux)
  - Make intuitive choices (e.g. red – warmer, blue – cooler; stronger colours for extreme values)
  - Avoid colours that are difficult to see or difficult to distinguish
  - How is missing data represented?
  - Consider how the figure would look in greyscale and to people with colour vision deficiencies

- Helpful design
  - Use sensible axis limits (optimise for single plot, or group of plots if comparison is important)
  - Label axes at sensible intervals and use tick marks
  - Label axes and colour-bars with variables and units
  - Design the layout to encourage comparison (e.g. side by side/stacked vertically)

- Be consistent with names, colours, markers (and style) throughout the report



Schultz (2009)



CONSIDER THE CONTEXT AND LIMITATIONS OF THE EVALUATION

- What are the main limitations of the evaluation?

- How do the results of this evaluation compare to previous findings (in other locations, with other models, with other model versions) and why?

- To what extent might the findings be transferrable?
  - How might the model perform in another city with different climate, building practices, energy use, socio-economic situation, human behaviour…
  - Can this be supported with sensitivity tests?
  - What situations would require further evaluation?

- What would be needed to improve model performance?