Model Validation: Comparing WRF with Observations

Lecture 8

Ph.D., Senior lecturer Madina Tursumbayeva

Department of Meteorology and Hydrology

Lecture Objective and Learning Outcomes

Lecture objective

To explain how to validate WRF output against observations: matching model and observations, statistical metrics and the interpretation of the differences.

Learning outcomes

By the end of this lecture, students will be able to:

1. explain why validation is necessary and name its four steps;
2. match model output with observations in space, height, time and variable;
3. calculate and interpret MBE, RMSE, the correlation coefficient and IOA;
4. explain why a single metric is not enough and use the daily cycle in validation;
5. name the main reasons why models and observations differ.

Why Validate a Model?

- A model almost always produces output, even when the set-up is wrong
- A wrong result looks as convincing as a correct one
- Validation: compare with independent observations and measure the difference

Not an examination

No model is perfect; the aim is to learn where the model is reliable and where it is not

Fit for purpose

A model can be good enough for one question and not good enough for another

The Four Steps of Validation

1. Observations

Independent of the data that drive the model

→

2. Matching

Same place, time, variable and units

→

3. Metrics

A few numbers that summarise all the pairs

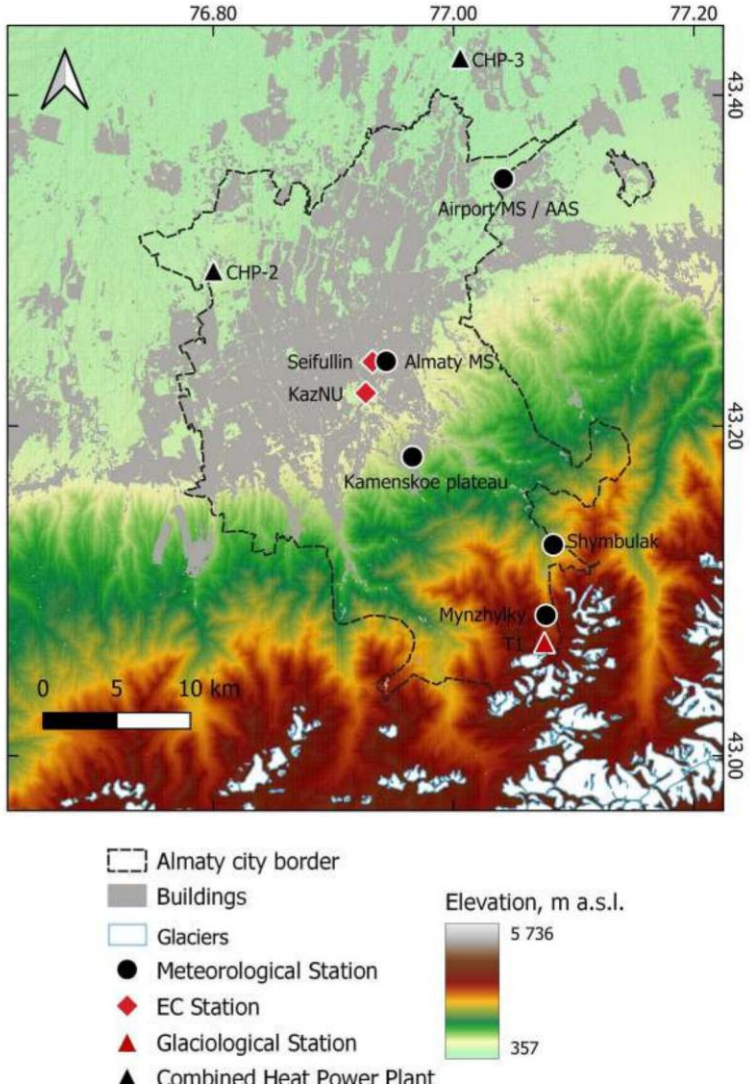
→

4. Interpretation

Why do they differ? What does it mean?

- Most mistakes are made in step 2
- A number alone means nothing without step 4
- Example throughout this lecture: our WRF-CHIMERE study of Almaty

Observations



- Surface stations: 2-m temperature, 10-m wind
- Radiosondes: vertical profiles, twice a day
- Air quality measurements: concentrations
- Satellites: spatial distribution

In our study

- 5 meteorological stations from 673 to 3017 m
- Radiosondes at the aerological station (AAS)
- Elemental carbon at 2 urban sites

Source: Tursumbayeva et al. (2026), Meteorological Applications, 33, e70183

Matching Model and Observations

Space

The nearest grid point: ll_to_xy(nc, lat, lon)

Height

Model terrain is smoothed: check the altitude of the grid point (within 30 m for 4 of our 5 stations)

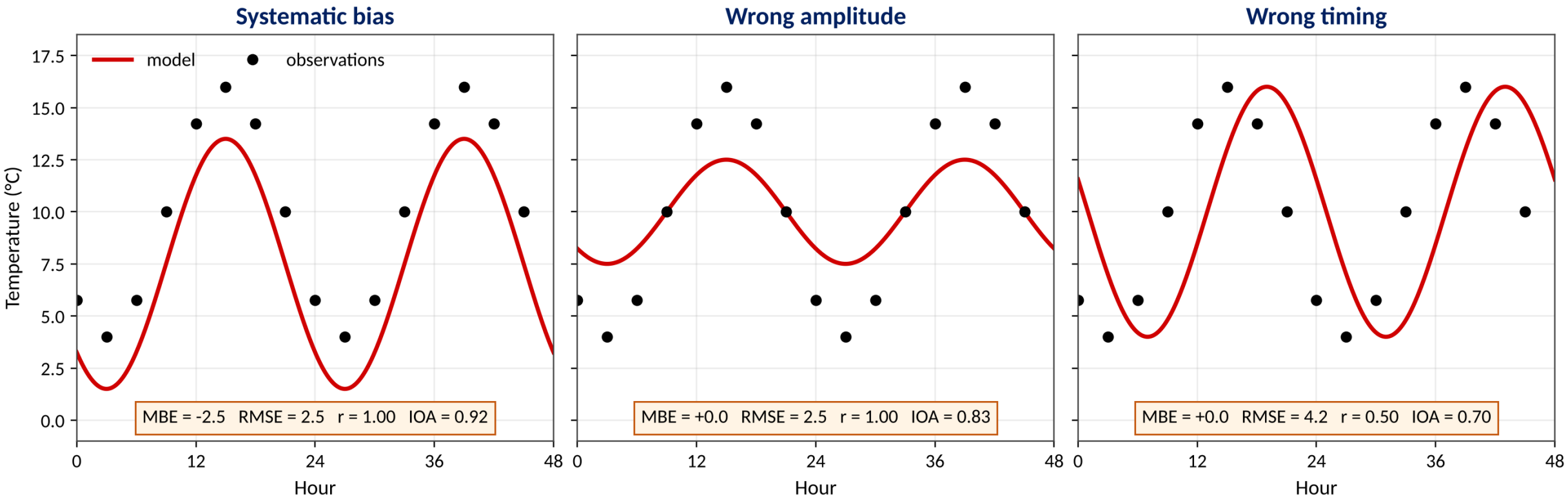
Time

Model output is in UTC; observations every 3 hours, often in local time

Variable

2-m temperature with 2-m temperature; 10-m wind rotated to true north; same units

Three Kinds of Error

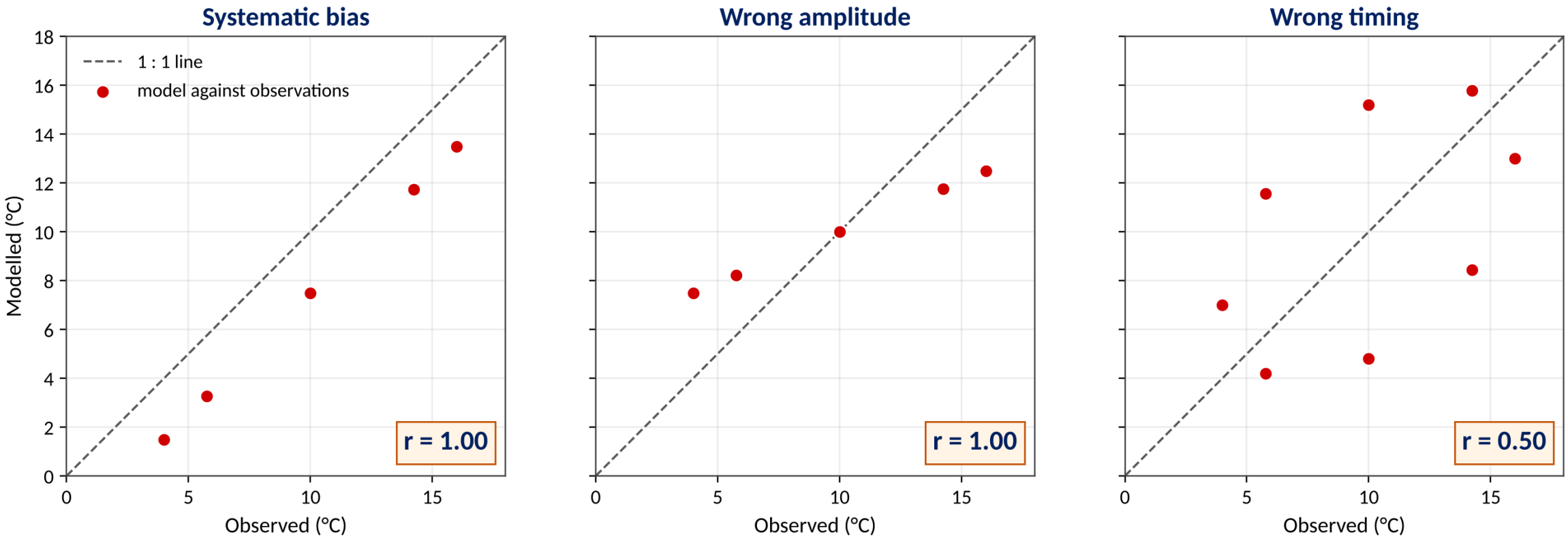


A zero mean bias does not mean a good model; equal RMSE does not mean equal errors

- One metric is never enough

Schematic example; the metrics are calculated from the plotted points

Correlation and the Scatter Plot



r = 1 for a biased model and for a wrong amplitude; only the timing error lowers it

- Correlation is sensitive to timing and pattern, but blind to bias and amplitude

Schematic example: the same three cases as on the previous slide

Four Metrics

Mean bias error (MBE)

$$MBE = mean(mod - obs)$$

Systematic error and its sign; ideal 0; errors of opposite sign cancel

Root mean square error (RMSE)

$$RMSE = \sqrt{mean((mod - obs)^2)}$$

Typical size of the error, in the units of the variable; ideal 0

Correlation coefficient (r)

$$r = cov(mod, obs) / (\sigma_{mod} \cdot \sigma_{obs})$$

From −1 to 1; do they vary together? Blind to bias and amplitude; ideal 1

Index of agreement (IOA)

$$IOA = 1 - \Sigma(mod - obs)^2 / \Sigma(|mod - \bar{o}bs| + |obs - \bar{o}bs|)^2$$

From 0 to 1; reduced also by bias and amplitude errors; ideal 1 (Willmott, 1981)

In our Almaty study: MBE, RMSE and IOA. Always report several metrics together

Code: Metrics in Python

```
import numpy as np

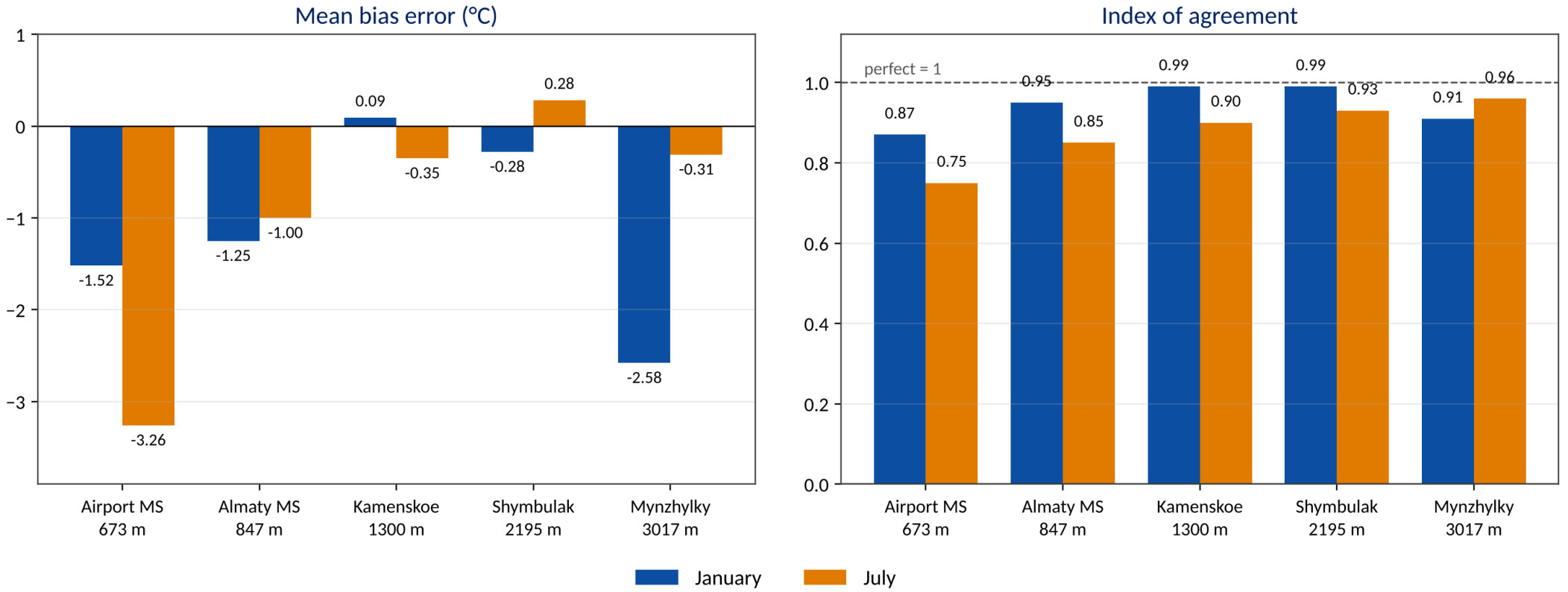
def metrics(mod, obs):
    ok = ~np.isnan(obs)              # skip missing data
    mod, obs = mod[ok], obs[ok]

    mbe  = np.mean(mod - obs)
    rmse = np.sqrt(np.mean((mod - obs)**2))
    r    = np.corrcoef(mod, obs)[0, 1]

    om  = obs.mean()
    ioa = 1 - np.sum((mod - obs)**2) / np.sum(
          (np.abs(mod - om) + np.abs(obs - om))**2)
    return mbe, rmse, r, ioa
```

- Two arrays of equal length, matched in time
- Remove the times with missing observations first
- Calculate separately for each station, variable and season

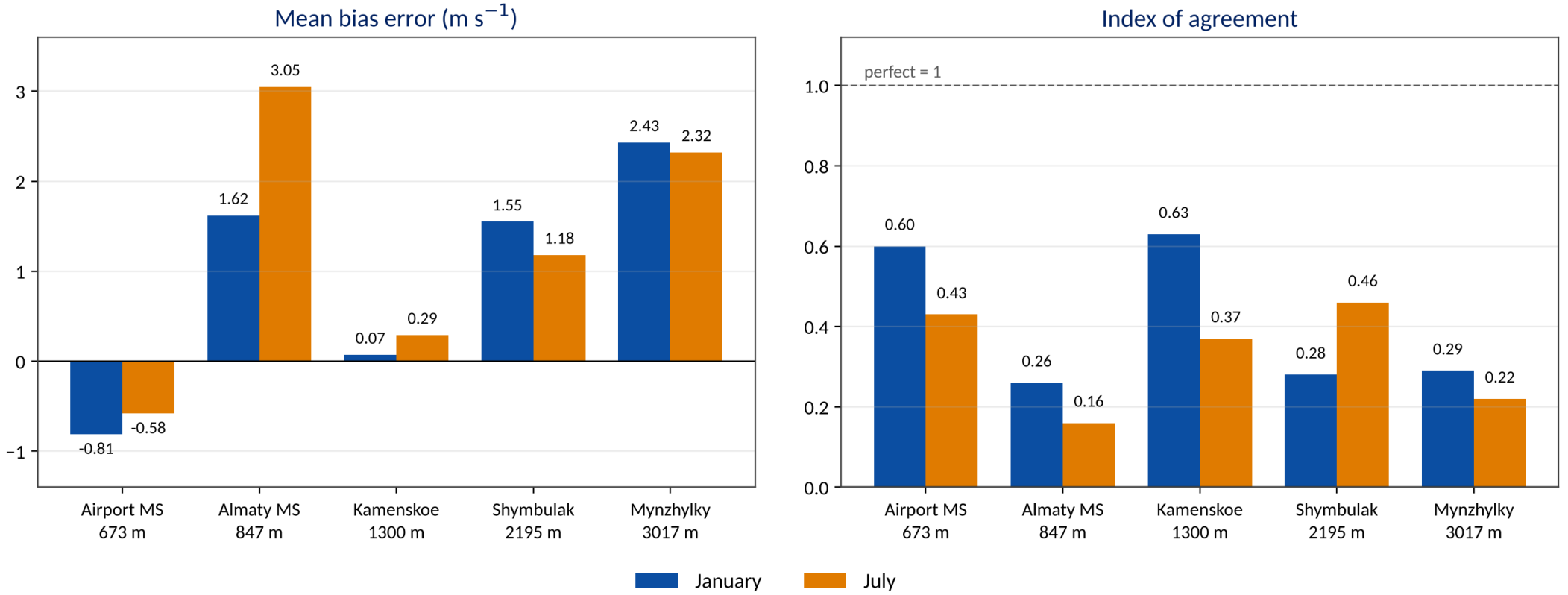
Research Example: 2-m Temperature



- IOA 0.75–0.99: the variations are reproduced well. Bias mostly negative: the model is too cold
- Largest bias at the airport (−3.3 °C in July) and at Mynzhylky (−2.6 °C in January): quality depends on place and season

Data: Table 2 of Tursumbayeva et al. (2026), Meteorological Applications, 33, e70183

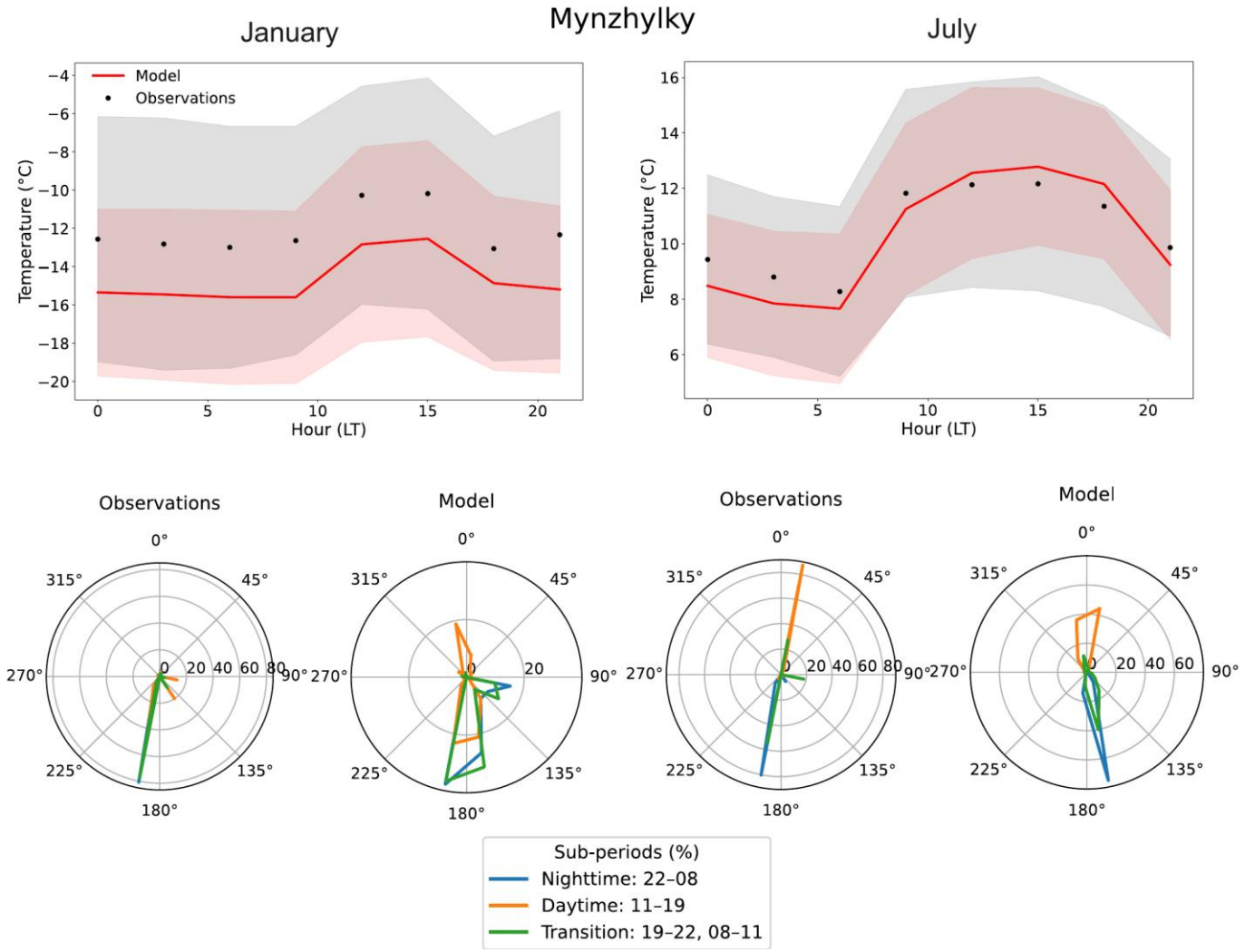
Research Example: 10-m Wind Speed



- Wind speed is overestimated everywhere except at the airport; IOA is much lower than for temperature
- Best in open terrain, lowest in the city centre (0.26 and 0.16): buildings and narrow valleys are smaller than a grid cell

Data: Table 2 of Tursumbayeva et al. (2026), Meteorological Applications, 33, e70183

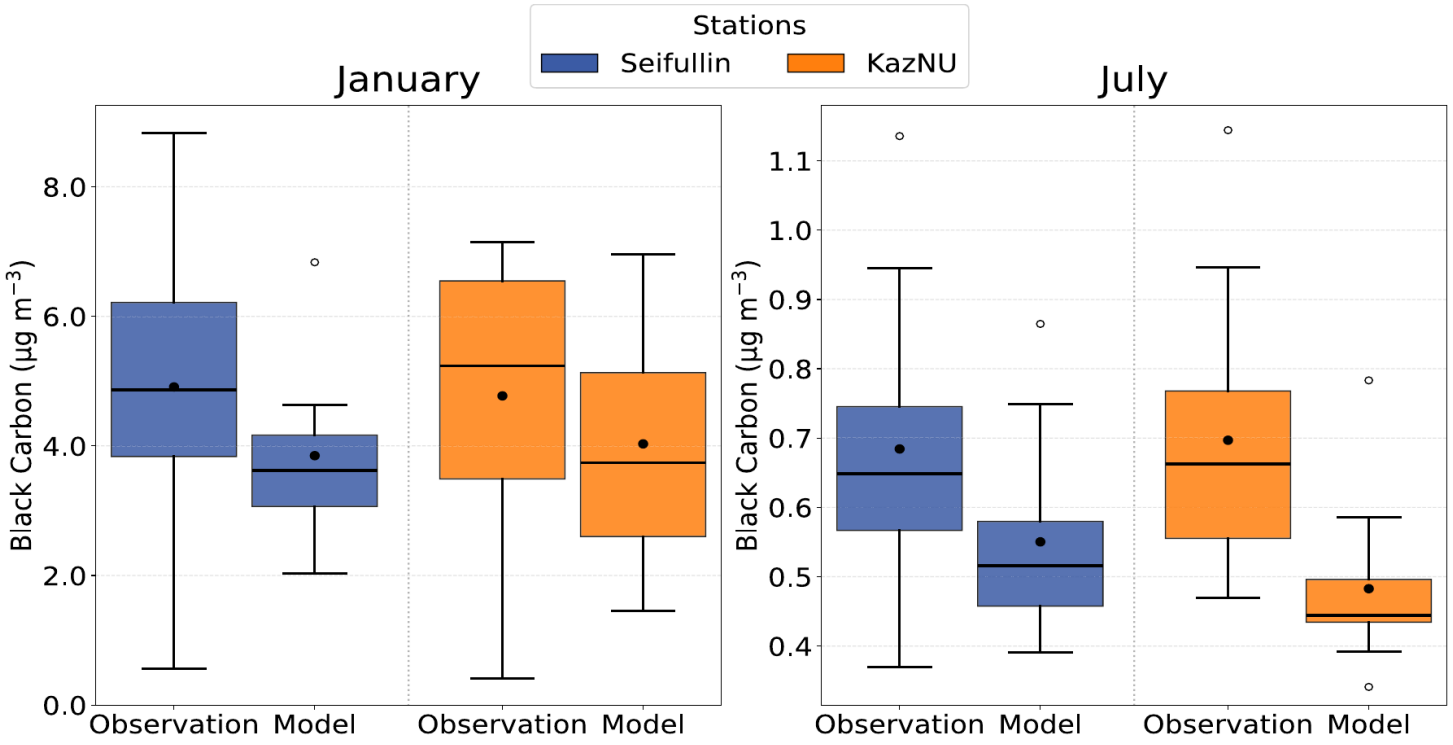
Beyond One Number: the Daily Cycle



- Mynzhylky, 3017 m: model (red), observations (black)
- July: the model follows the observations
- January: correct shape, but colder at all hours
- Wind roses for night, day and transition hours: the daily reversal of the valley wind is captured
- Validate processes, not only values

Source: Tursumbayeva et al. (2026), Meteorological Applications, 33, e70183

Validating the Pollutant: Black Carbon

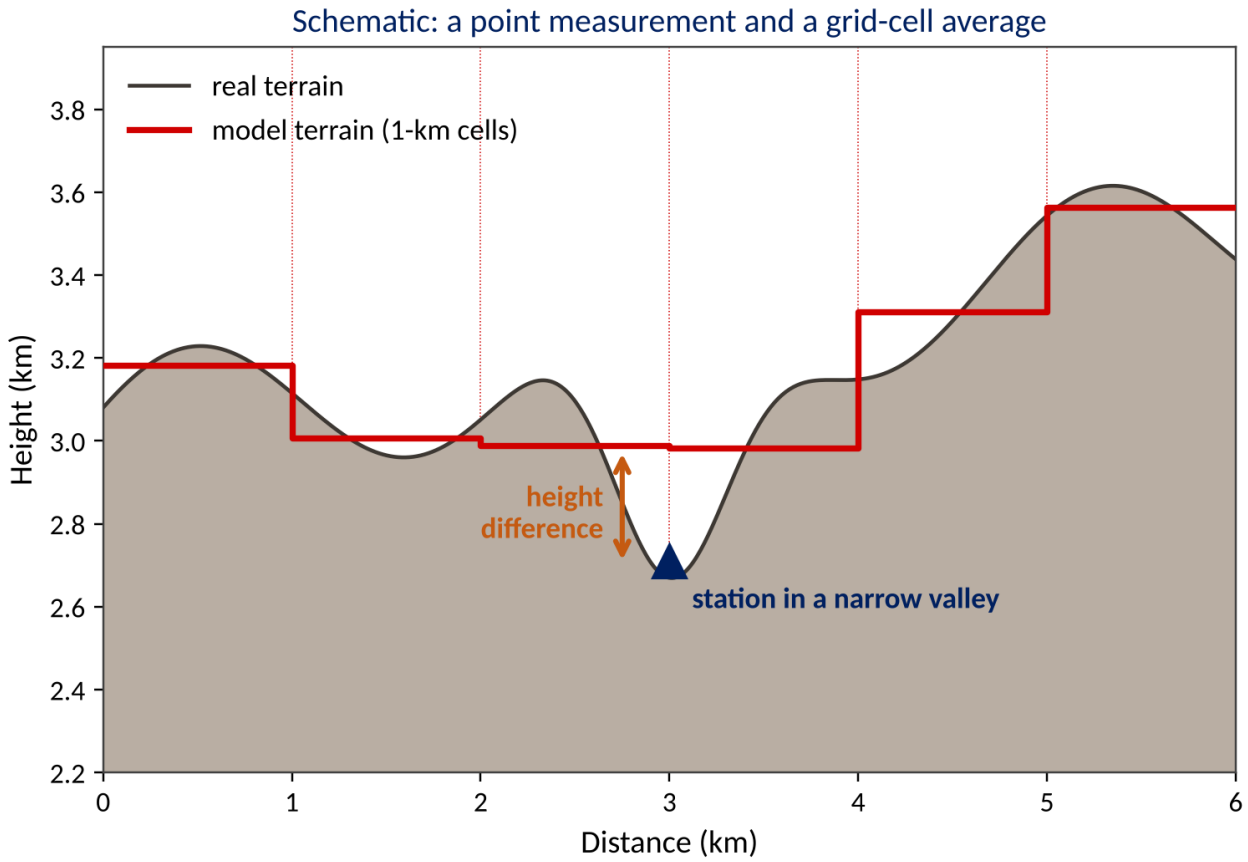


- Seasonal contrast reproduced: about 4–5 μg/m³ in January, below 1 in July
- The model is somewhat lower than the observations
- IOA: 0.4–0.6 in winter, about 0.3 in summer
- Errors of emissions, wind and mixing add up

Compare like with like: the model simulates black carbon, the instrument measures elemental carbon

Source: Tursumbayeva et al. (2026), Meteorological Applications, 33, e70183

Why Do Models and Observations Differ?



- Representativeness: a point against a grid-cell average
- Input data: boundary conditions, land use, emissions
- Model physics: boundary layer, surface, radiation, clouds
- Resolution: buildings and narrow valleys are sub-grid
- Observations: instrument errors, local effects

A difference is not always an error of the model

Checklist for a Good Validation

Several variables

Temperature, wind and the quantity of interest

Several stations

Different parts of the domain and altitudes

Several metrics

MBE, RMSE, correlation and IOA

Daily cycle and seasons

Not only one number for the whole period

The vertical

Profiles, if soundings are available

Honest reporting

State the weaknesses of the model

Summary

Key points

- Validation: comparison with independent observations
- Match in space, height, time and variable
- MBE: systematic error. RMSE: size of the error. r: common variation. IOA: overall agreement
- No single number is enough: look at the daily cycle
- The aim: to know where the model can be trusted

Next lecture

Scenario analysis

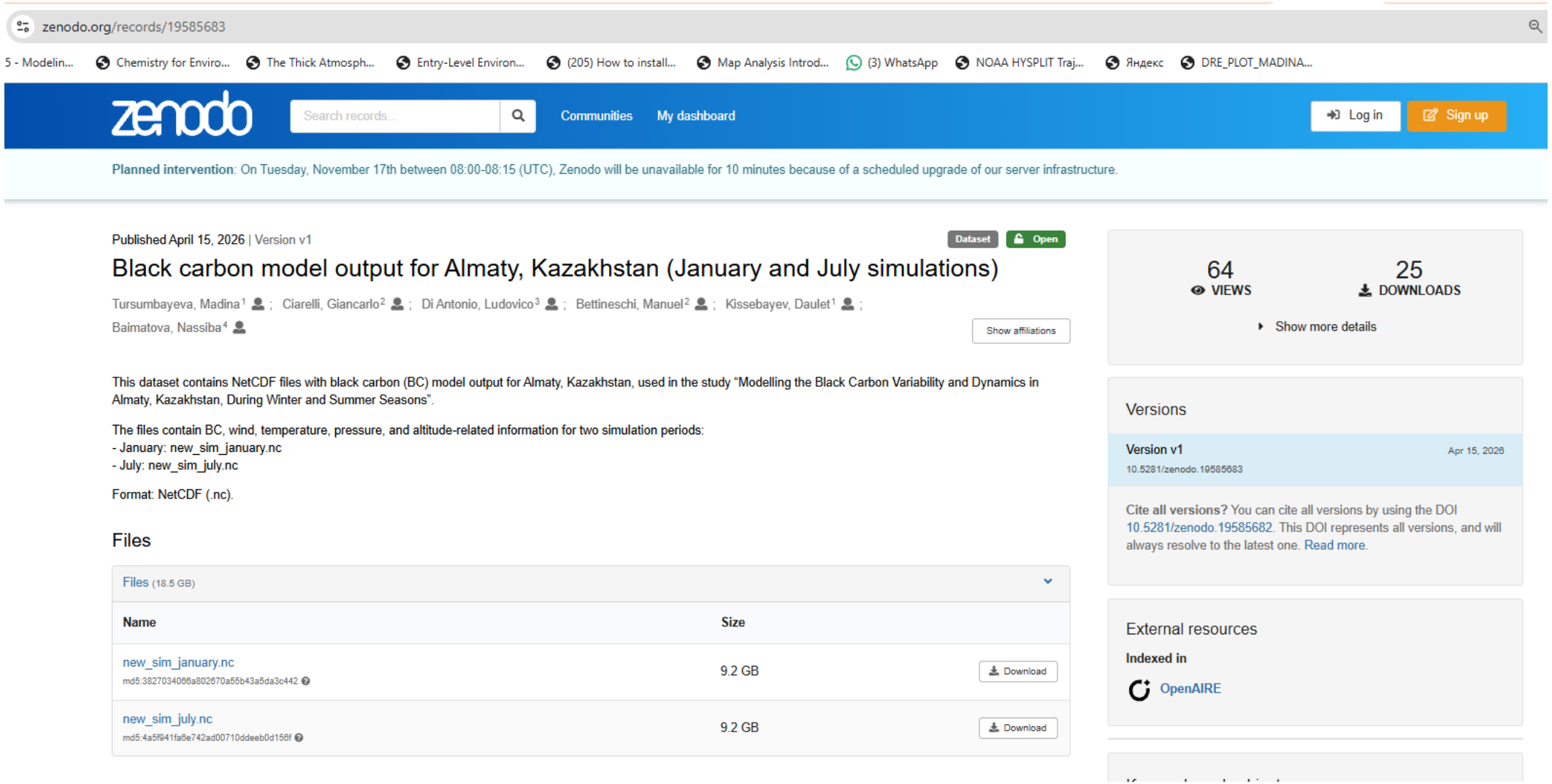
Experiments with the model: what would happen if…?

Thank you for your attention!

Self-Check Questions

1. Why must a model be validated even when the run has finished without errors?
2. What are the four steps of validation, and at which step are most mistakes made?
3. What must be checked when a station is matched with a model grid point?
4. What does each of MBE, RMSE, r and IOA show? Which of them is blind to bias?
5. Why is wind speed in a city reproduced worse than temperature?
6. Why is a difference between model and observations not always an error of the model?

Practical work: Zenodo database



References

1. Tursumbayeva, M., Ciarelli, G., Di Antonio, L., Bettineschi, M., Kissebayev, D., & Baimatova, N. (2026). Modelling the Black Carbon Variability and Dynamics in Almaty, Kazakhstan, During Winter and Summer Seasons. Meteorological Applications, 33, e70183. https://doi.org/10.1002/met.70183
2. Willmott, C. J. (1981). On the validation of models. Physical Geography, 2(2), 184–194.
3. Wilks, D. S. (2011). Statistical Methods in the Atmospheric Sciences (3rd ed.). Academic Press.
4. NSF NCAR Mesoscale and Microscale Meteorology Laboratory. WRF Tutorial: recorded lectures (YouTube, wrfhelp NCAR-MMM, 2021) and lecture slides, including “Verification of WRF Simulations”. https://www2.mmm.ucar.edu/wrf/wrf_tutorial/lectures.html
5. Skamarock, W. C., et al. (2019). A Description of the Advanced Research WRF Model Version 4. NCAR Technical Note NCAR/TN-556+STR. https://doi.org/10.5065/1dfh-6p97
6. Ladwig, W. (2017). wrf-python (Version 1.3) [Software]. UCAR/NCAR. https://doi.org/10.5065/D6W094P1