Evaluating generated in situ weather forecasts with exchangeability methods

David Landry, Anastase Charantonis, Claire Monteleoni
INRIA Paris

Ver-AI workshop on Verification
Reading
2025-06-23

Overview

Case study where multi-variate rank histogram methods helped in the evaluation and calibration of generative weather forecasting model.

1. Introduce an in situ generative postprocessing model
2. Diagnose a forecast spread issue using multivariate rank histograms
3. Calibrate the generative model

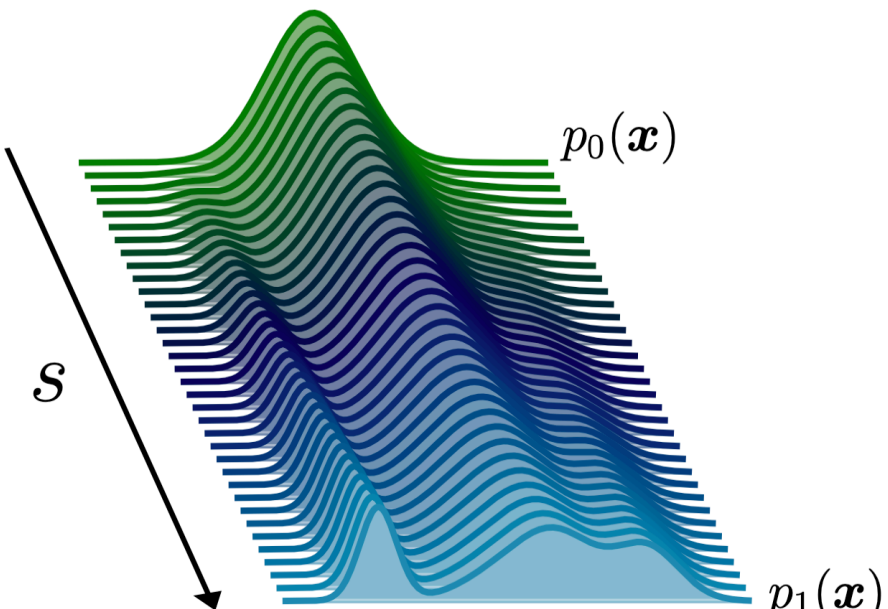
In situ forecasting, but the methodology should apply to gridded forecasts as well.

Generative in situ forecasting model

Flow matching postprocessing (FMAP)

Generative in situ postprocessing method with a particular focus modeling on **spatial correlations** (Landry et al. 2025).

Based on **flow matching** (very closely related to **denoising diffusion methods**).



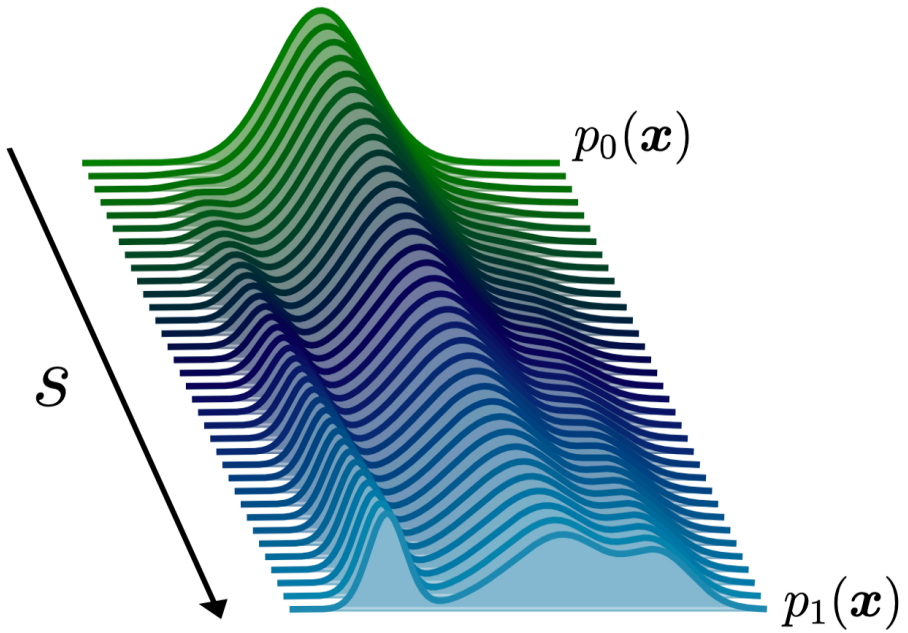
Key characteristics

The statistical objects the generative model approximates are **samples** (as opposed to the conditional expectation).

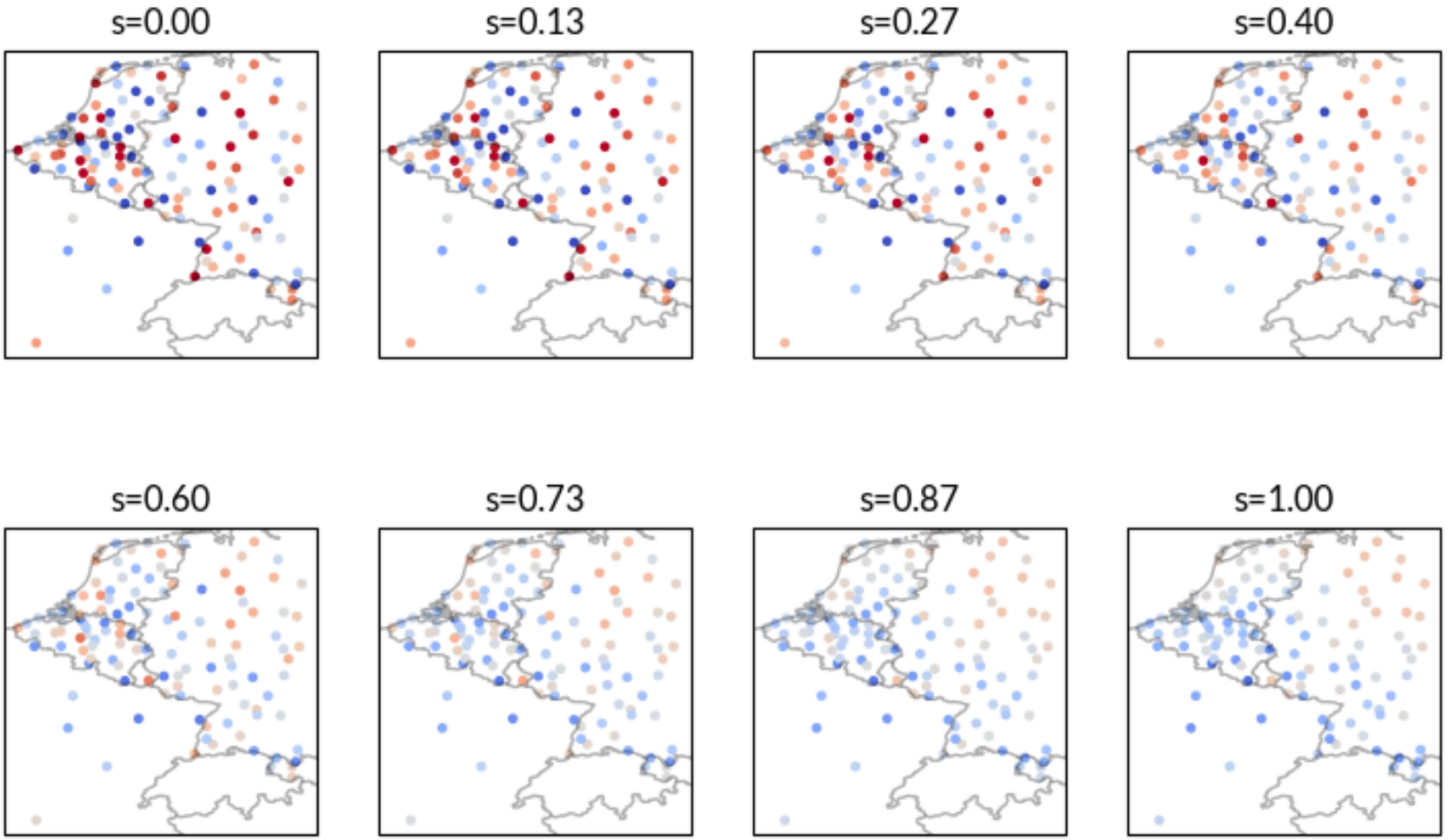
A numerical integration displaces the samples from a normal distirbution to the target distribution.

$$p_0(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

$$p_1(\boldsymbol{x}) \approx \text{target distribution}$$



Flow matching distribution transport



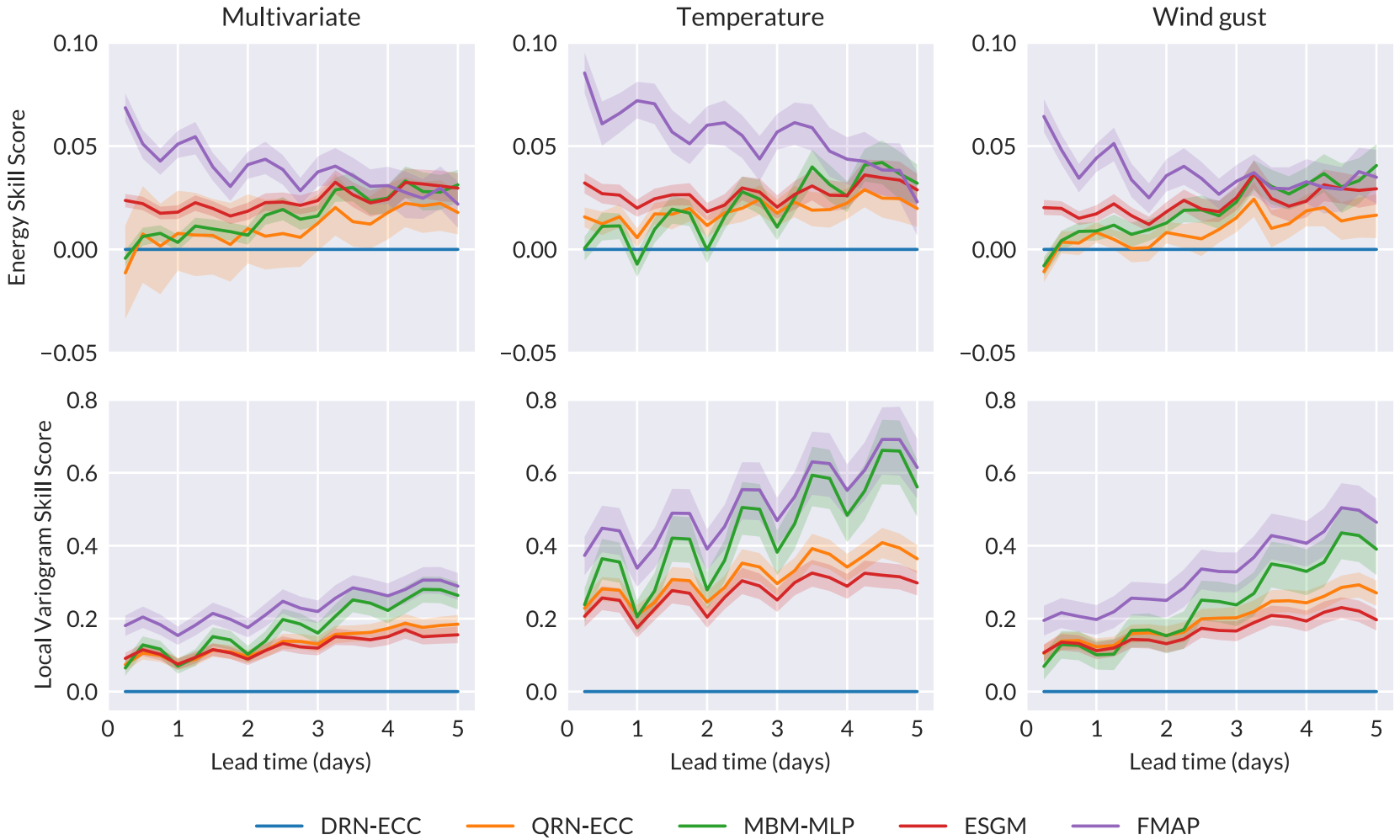
Example flow integration

Experiments

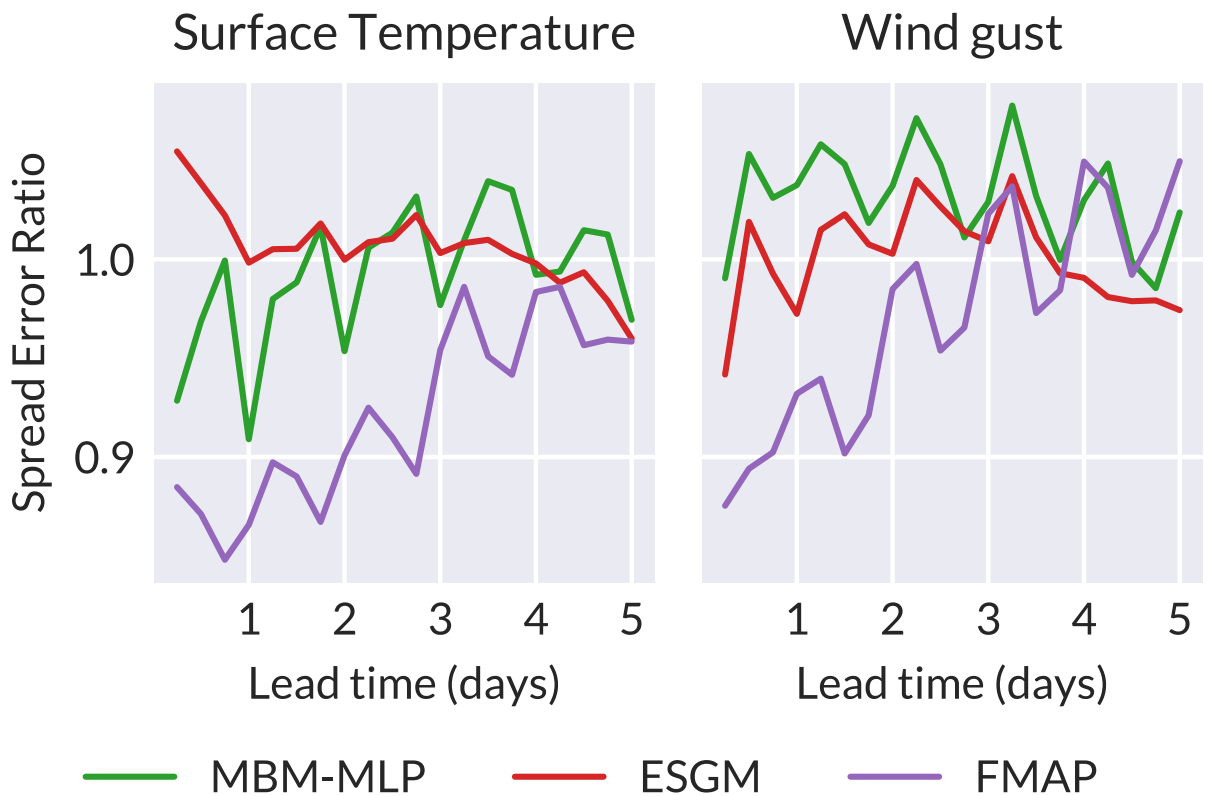
Dataset EUPPBench (Demaeyer et al. 2023)

- 0.25° resolution
- 20 years reforecasts (bi-weekly)
- 2 years forecasts (daily)

Target: Surface temperature and wind gust at 121 stations in central Europe.



Flow matching model skill score



Spread-error ratio for flow matching model (FMAP) and other baselines

Dispersion behavior of generative models

Both **GenCast** (Price et al. 2025) and **ArchesWeatherGen** (Couairon et al. 2024) are under-dispersive.

The latter proposes a calibration procedure to compensate.

Multivariate rank histograms to evaluate dispersion

Evaluation strategy

We calibrate the well-known rank histograms, but in a multivariate setting.

Our multivariate vector x is the forecast for one variable at every station (121-long).

Evaluation strategy

We will use methods based on

- **Minimum Spanning Tree (MST)** (Smith and Hansen 2004; Wilks 2004)
- **The Mahalanobis distance**

These methods give us some insight on the dispersion behavior of the models.

Exchangeability methods for evaluation

We assume the ensemble

$$\{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$$

is **exchangeable** where $\boldsymbol{x}_0$ is the verifying observation and $\boldsymbol{x}_{1...m}$ the ensemble members.

Informally, we verify that observation x_0 behaves like any other member.

Pre-rank functions

A straightforward strategy to build rank histograms for multivariate quantities is to use a **pre-rank function** (Gneiting et al. 2008).

$$F : \mathbb{R}^d \times \underbrace{\mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{\text{Symmetric}} \longrightarrow \mathbb{R}$$

The prerank of member j is

$$z_j = F\left(\boldsymbol{x}_j; \boldsymbol{x}_{-j}\right)$$

where $\boldsymbol{x}_{-j} = \{\boldsymbol{x}_{0..m}\} \setminus \{\boldsymbol{x}_j\}$.

Allowable pre-rank functions are **symmetric** on the right-hand side argument to preserve exchangeability (rank histogram flatness).

Minimum spanning tree

A proposal by (Smith and Hansen 2004; Wilks 2004)

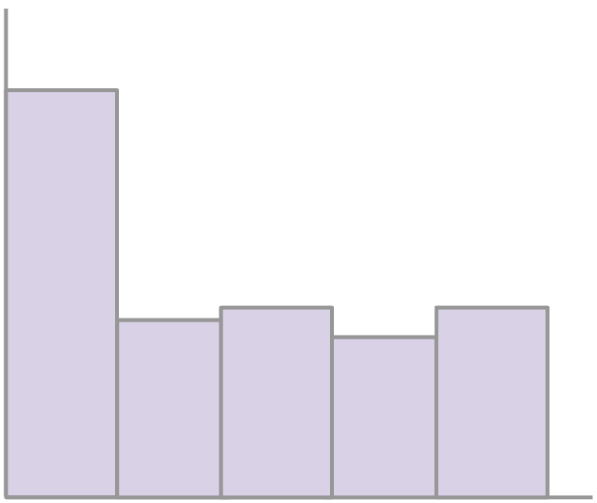
Build a fully-connected graph where

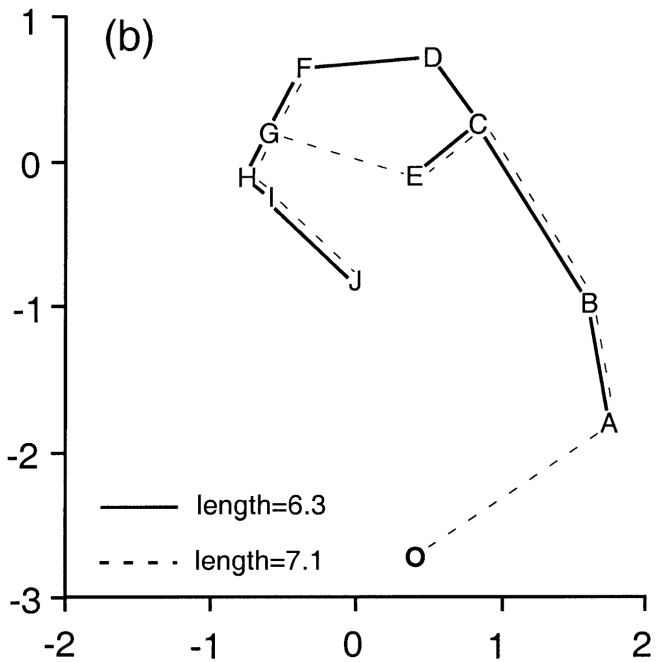
- the nodes are the ensemble members
- the edges have length $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$

From that graph extract the **minimum spanning tree** (MST) and measure its length.

$$F_{\mathrm{MST}(\boldsymbol{x}_j;\boldsymbol{x}_{-j})} = \mathrm{MST}\left(\boldsymbol{x}_{-j}\right)$$

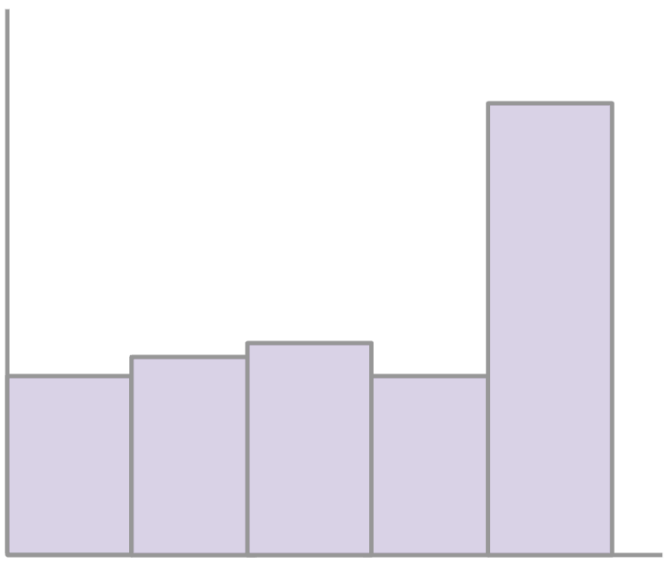
Interpreting MST histograms

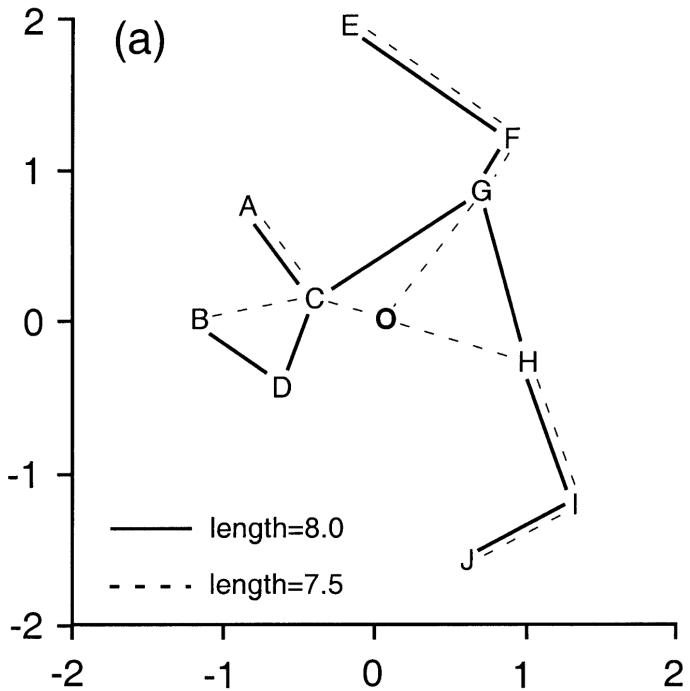




When the observation is removed, the MST is shorter. Underdispersion, systematic bias.

Interpreting MST histograms





When the observation is removed, the MST is longer. Overdispersion.

Mahalanobis distance

Sample to normal distribution metric.

Given a distribution $Q = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we have

$$\mathrm{MD}(Q, \boldsymbol{y}) = \sqrt{(\boldsymbol{\mu} - \boldsymbol{y})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{y})}.$$

Mahalanobis prerank function

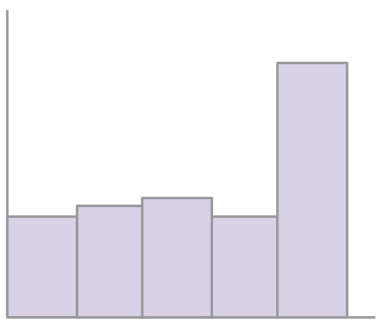
We propose

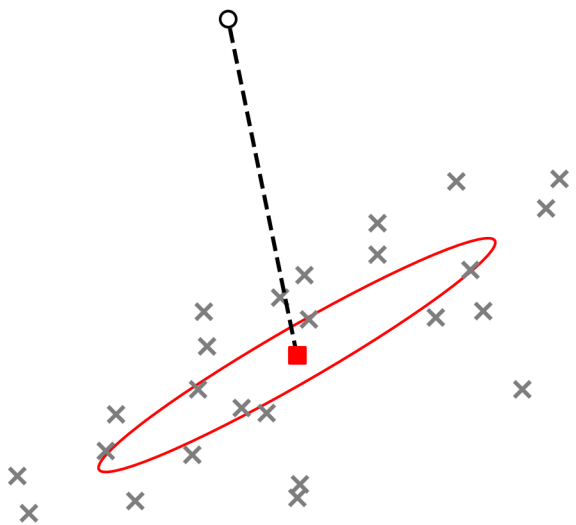
$$z_j = F_{\mathrm{MD}}\left(\boldsymbol{x}_j; \boldsymbol{x}_{-j}\right) = \mathrm{MD}\left(Q_{\boldsymbol{x}_{-j}}, \boldsymbol{x}_j\right)$$

where $Q_{x_{-j}}$ is estimated from the ensemble. *This may require specific estimation methods such as a Ledoit-Wolf shrinkage.*

Has a similar interpretation to the MST (in reverse).

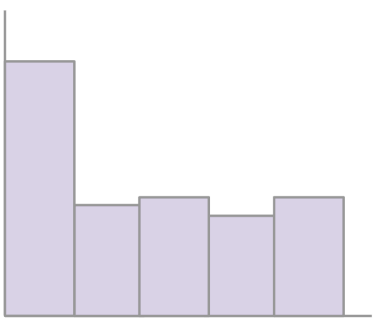
Interpreting MD histograms

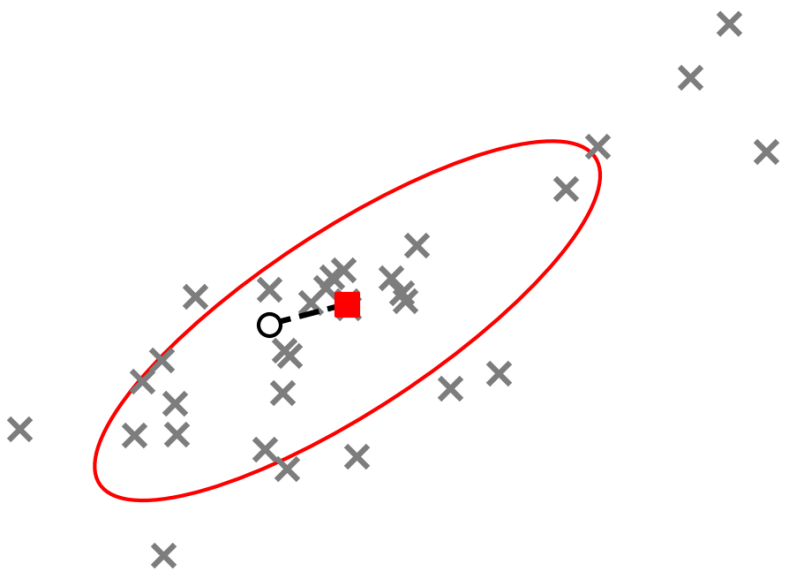




Observation is far from the ensemble distribution → Underdispersion, systematic biases.

Interpreting MD histograms





Observation is central within the ensemble → Overdispersion.

What about the Box Ordinate Transform (BOT)?

This proposed method is similar to the BOT evaluation method (Gneiting et al. 2008). Instead of a prerank it computes

$$u = 1 - \chi^2_d \left[(\boldsymbol{x}_0 - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}) \right]$$

which we translate to a bin number in a PIT histogram using

$$b = 1 + \lfloor u(m+1) \rfloor.$$

We get uniform u if the normality assumption is true.

What about the Box Ordinate Transform (BOT)?

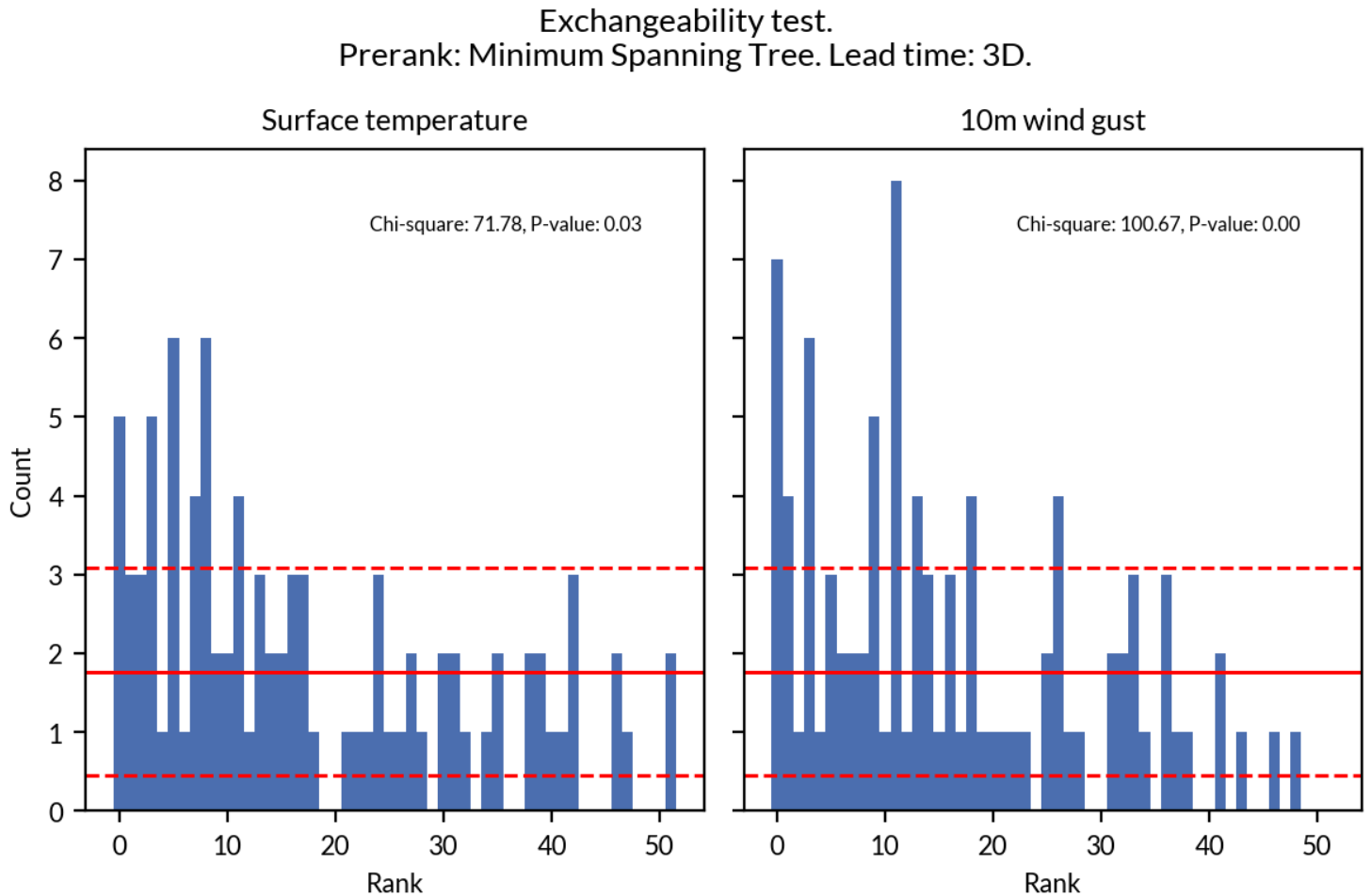
Our model fails the BOT test consistently

- Normal assumption is false – the χ^2 test saturates

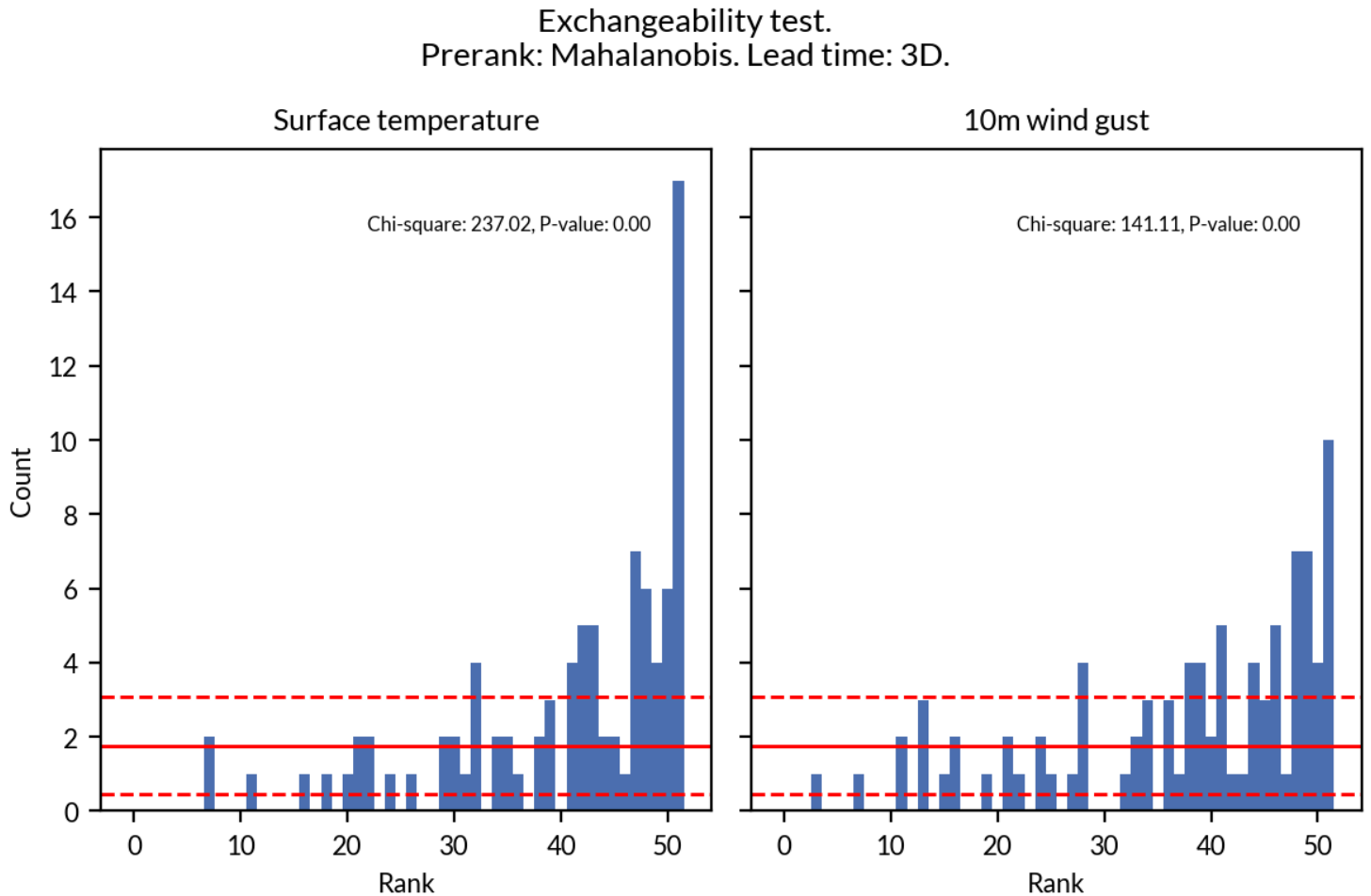
Our Mahalanobis pre-rank approach is less sensitive to the normality of the data. It ensures that all members are equally non-normal.

It requires m times more matrix inversions.

Experiments on EUPP postprocessing



MST test on 84 days calibration period



Mahalanobis distance test on 84 days calibration period

Calibrating the flow matching model

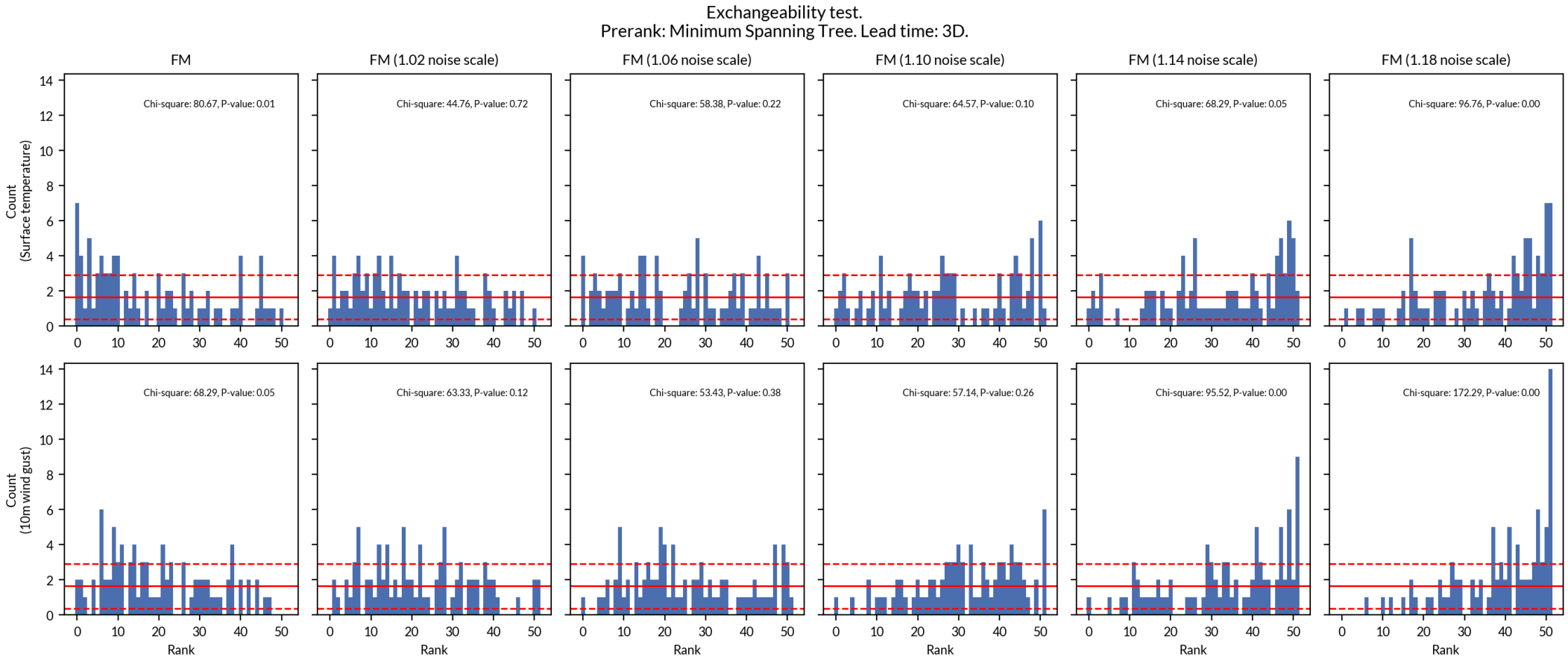
Calibrating the flow matching model

We propose to use the multivariate rank histograms to calibrate the flow matching model as in Couairon et al. (2024).

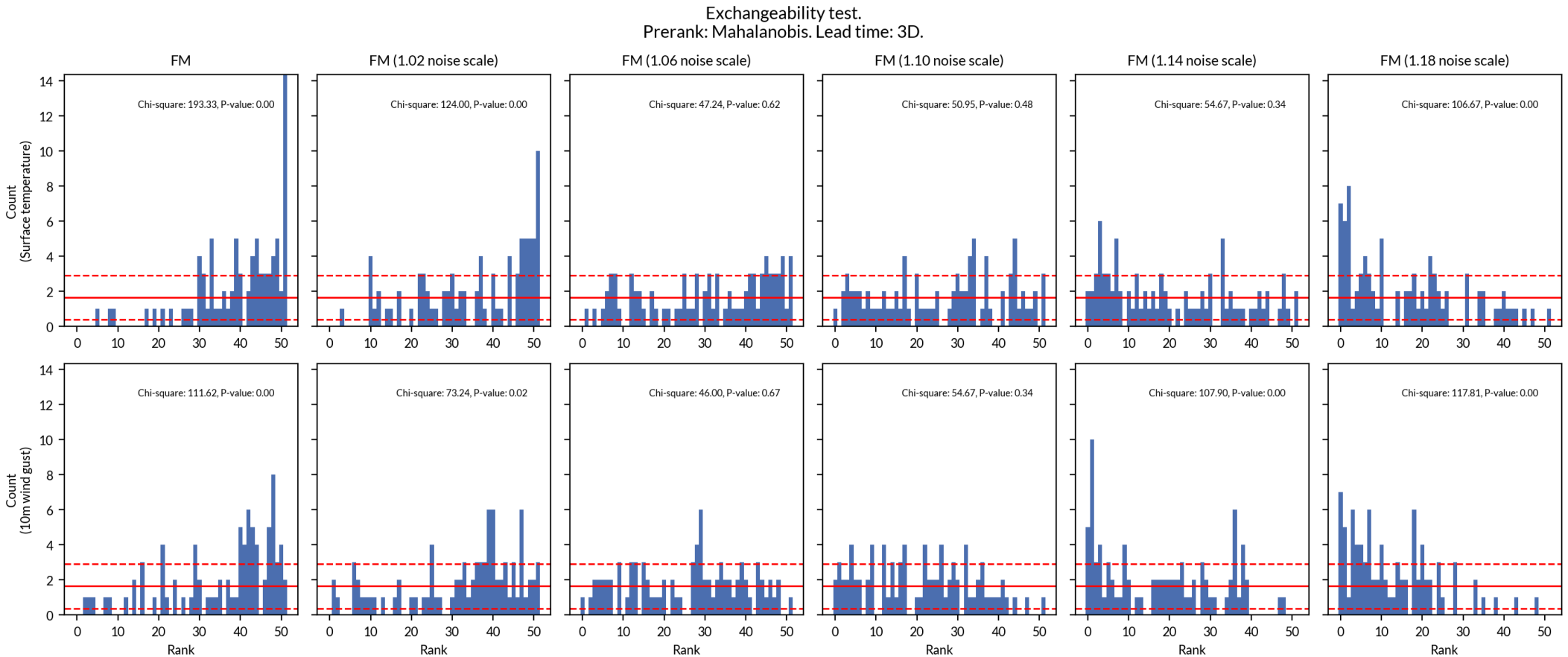
The initial sample is rescaled such that

$$p_0(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \alpha \boldsymbol{I}).$$

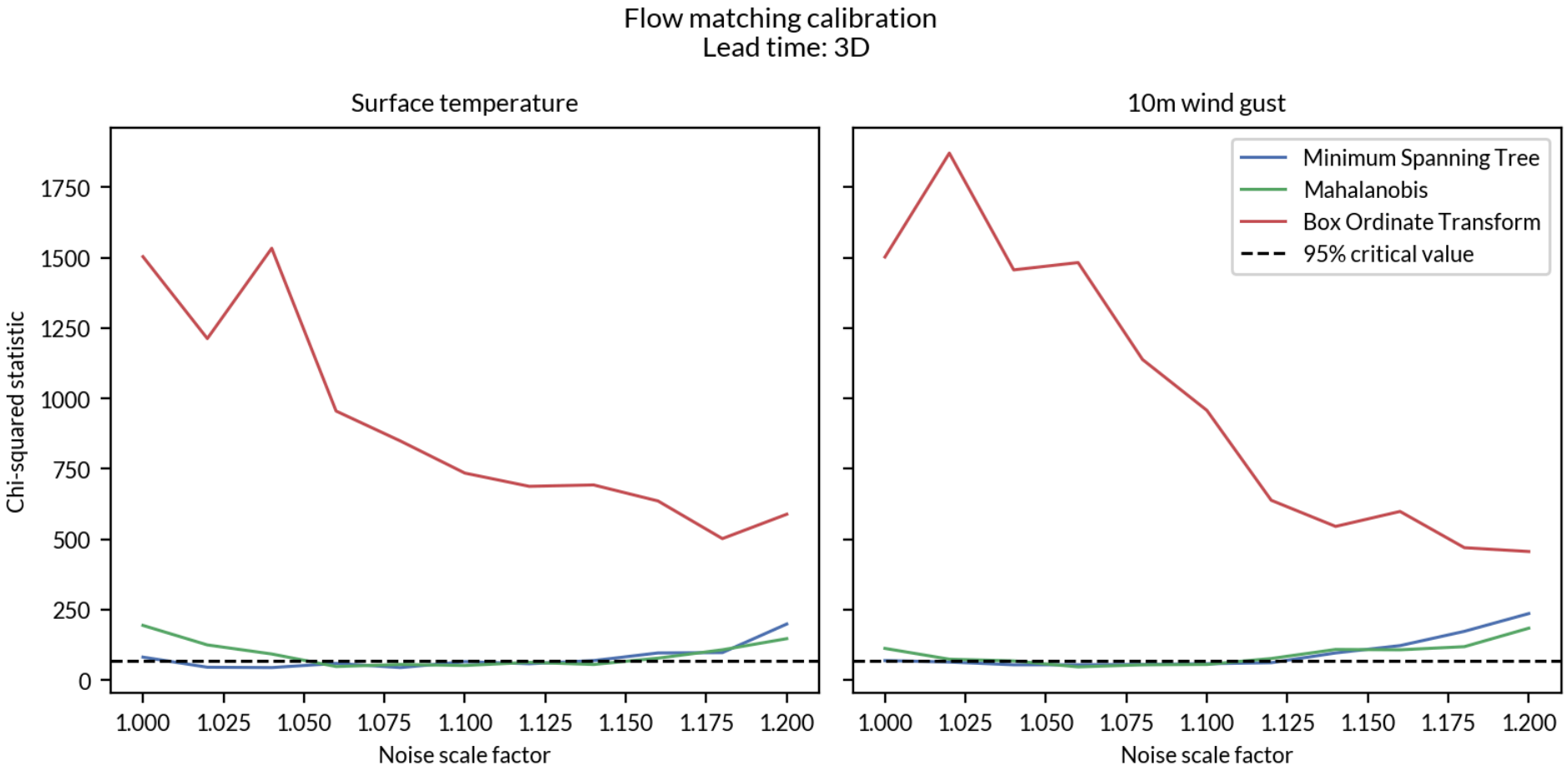
We reserved an 84 days calibration period (1st week of each month in 2017).



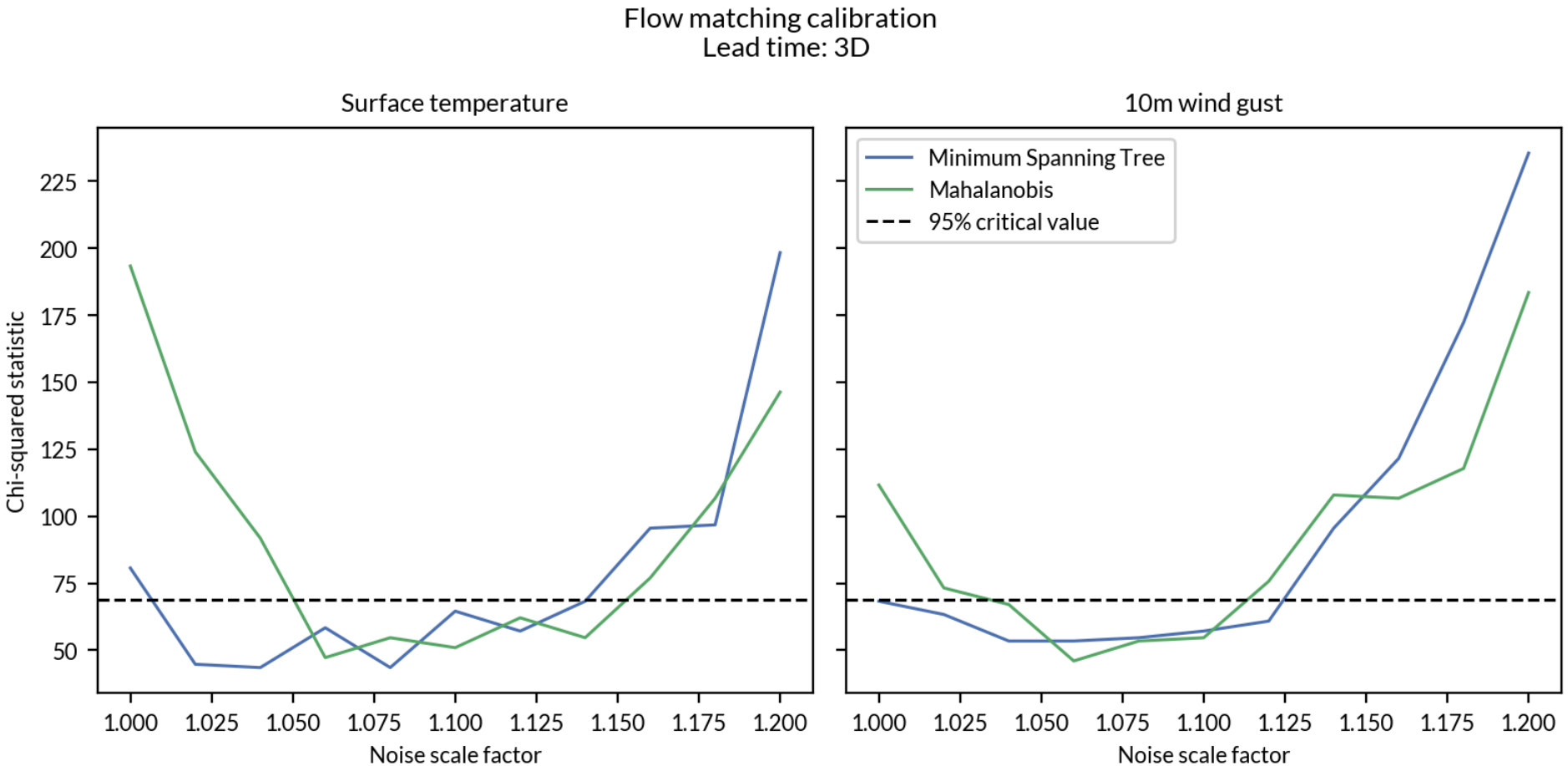
Rank histograms for flow matching model calibration (MST)



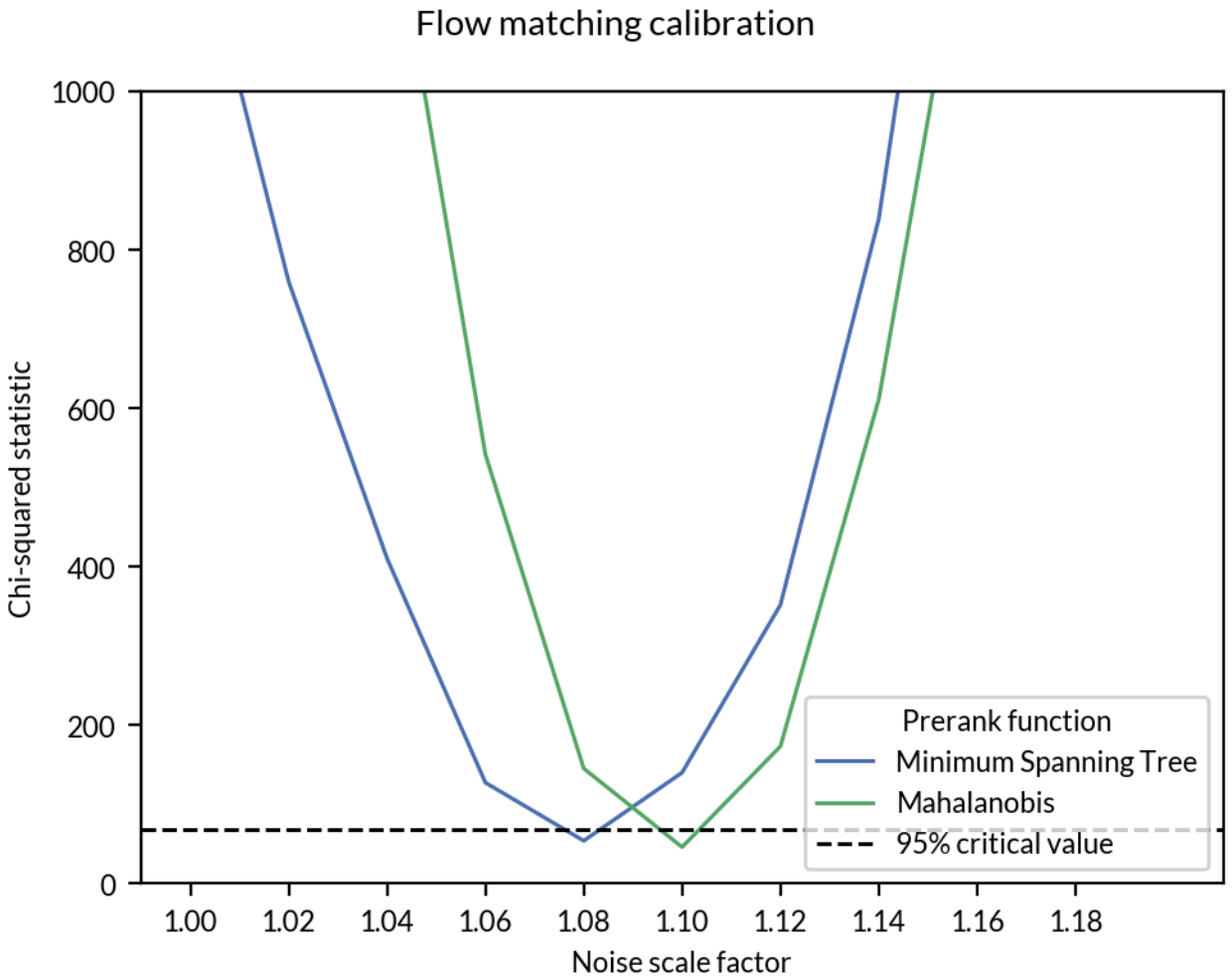
Rank histograms for flow matching model calibration (Mahalanobis)



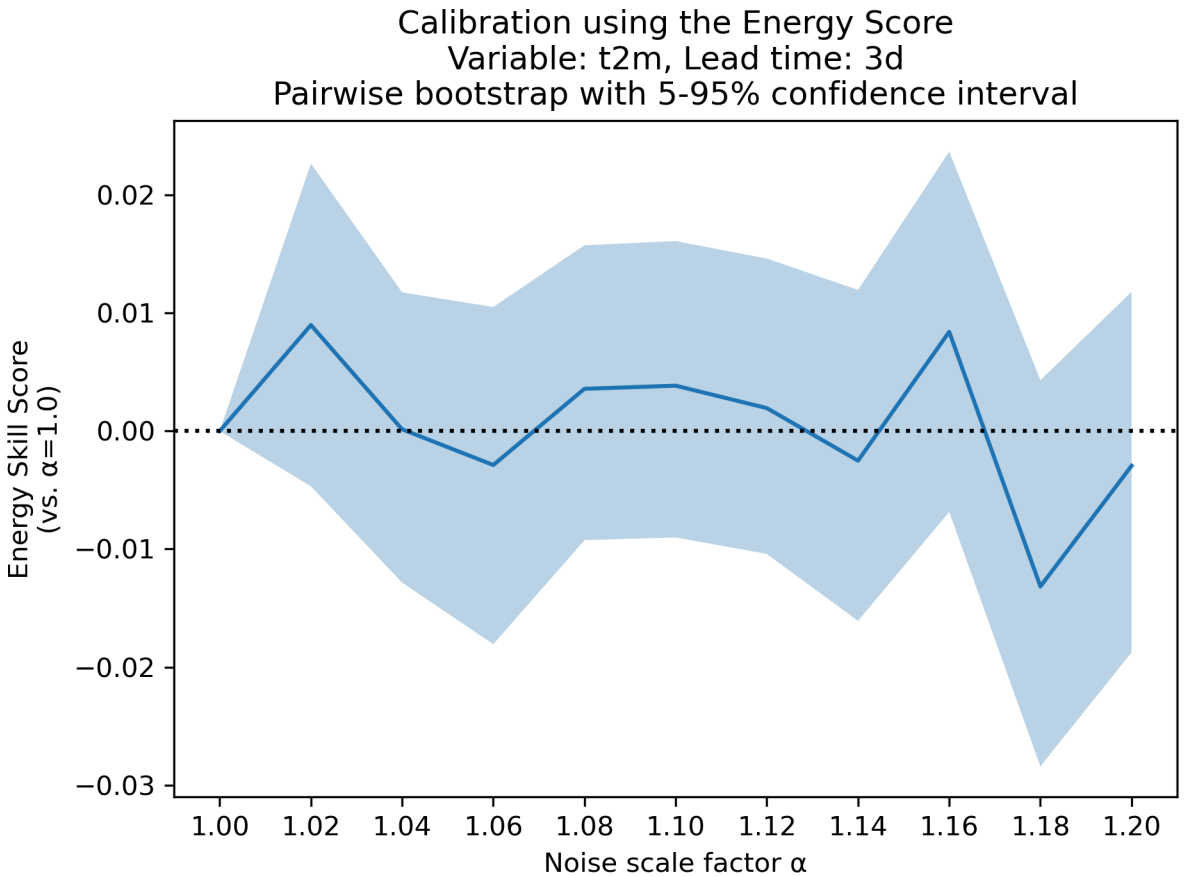
Chi-square test statistic given the scaling factor



Chi-square test statistic given the scaling factor

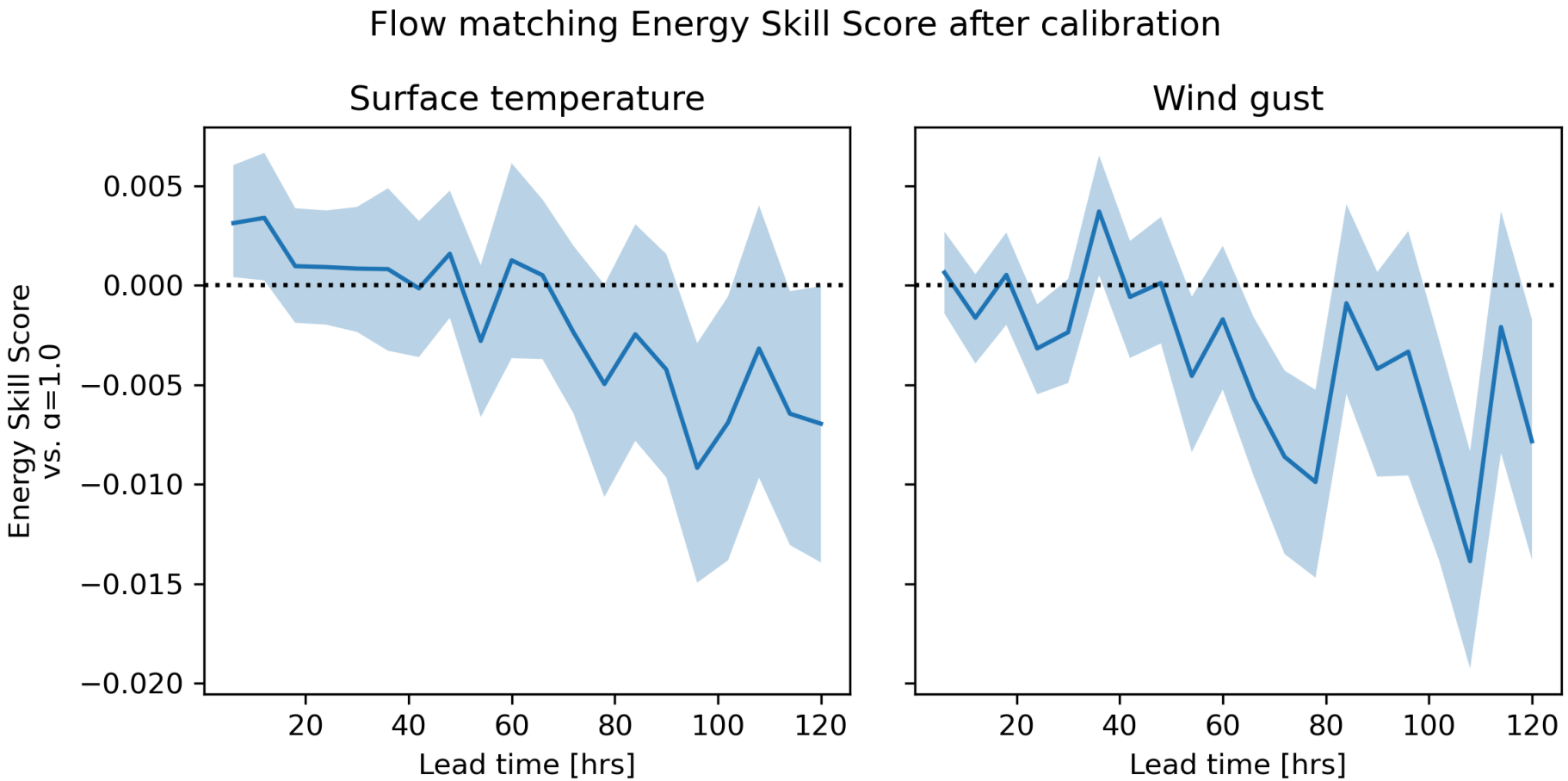


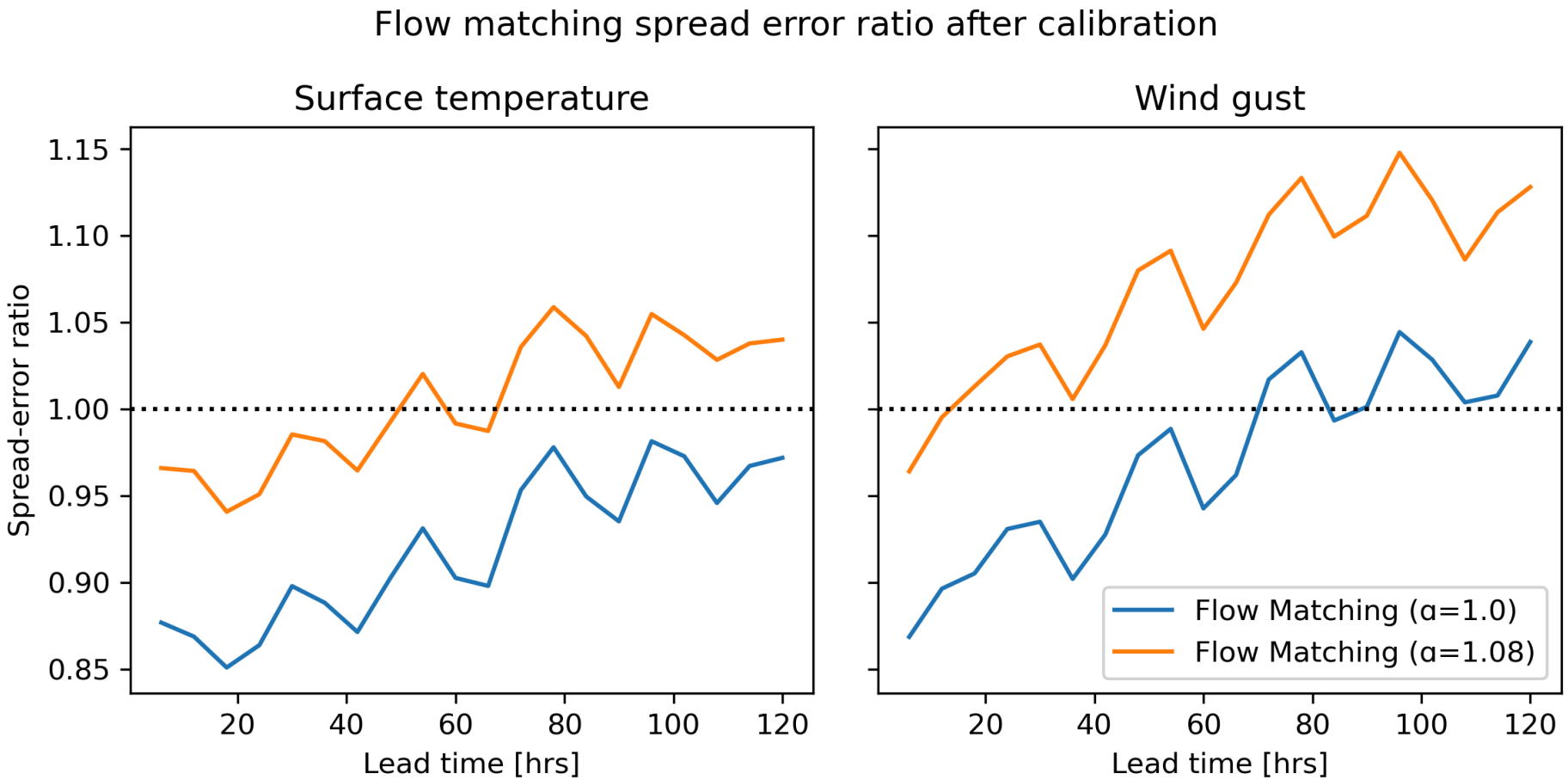
Stacked all lead-time-variable combinations



This calibration would require a larger calibration set using the ES

Effect of calibration on other metrics


Flow matching Energy Skill Score after calibration
Surface temperature
Wind gust
Energy Skill Score
vs. α=1.0
0.005
0.000
−0.005
−0.010
−0.015
−0.020
20
40
60
80
100
120
Lead time [hrs]


Flow matching spread error ratio after calibration
Surface temperature
Wind gust
Spread-error ratio
1.15
1.10
1.05
1.00
0.95
0.90
0.85
20
40
60
80
100
120
Lead time [hrs]
Flow Matching (α=1.0)
Flow Matching (α=1.08)

Discussion

The α parameter has an important impact on dispersion, as measured by out reliability tests.

Calibration seems successful for temperature, but less so on wind gust.

Any interactions with the normalization procedure on wind?

Conclusion

In summary...

- Multivariate rank histograms were useful in evaluating an in situ generative weather forecasting model
- We proposed an exchangeability test similar to BOT which is less sensitive to the normality of the data
- This allowed a multivariate calibration procedure would have been challenging on the Energy Score

Conclusion

In the future...

- How does this translate to higher dimensionalities like in full NWP forecasts?
- Better account for test multiplicity and serial correlations (Wilks 2019)

References

Couairon, G., R. Singh, A. Charantonis, C. Lessig, and C. Monteleoni, 2024: ArchesWeather & ArchesWeatherGen: A Deterministic and Generative Model for Efficient ML Weather Forecasting. https://doi.org/10.48550/arXiv.2412.12971.

Demaeyer, J., and Coauthors, 2023: The EUPPBench Postprocessing Benchmark Dataset v1.0. *Earth System Science Data Discussions,* 1–25, https://doi.org/10.5194/essd-2022-465.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson, 2008: Assessing Probabilistic Forecasts of Multivariate Quantities, with an Application to Ensemble Predictions of Surface Winds. *TEST,* **17**, 211–235, https://doi.org/10.1007/s11749-008-0114-x.

Landry, D., C. Monteleoni, and A. Charantonis, 2025: Generating Ensembles of Spatially-Coherent in-Situ Forecasts Using Flow Matching. https://doi.org/10.48550/arXiv.2504.03463.

Price, I., and Coauthors, 2025: Probabilistic Weather Forecasting with Machine Learning. *Nature,* **637**, 84–90, https://doi.org/10.1038/s41586-024-08252-9.

Smith, L. A., and J. A. Hansen, 2004: Extending the Limits of Ensemble Forecast Verification with the Minimum Spanning Tree. *Monthly Weather Review,* **132**, 1522–1528.

Wilks, D. S., 2004: The Minimum Spanning Tree Histogram as a Verification Tool for Multidimensional Ensemble Forecasts. *Monthly Weather Review,* **132**, 1329–1340.

Wilks, D. S., 2019: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 4th ed. Elsevier,.

Thank you

FMAP preprint

Email