Evaluating multivariate ensemble forecasts

Martin Leutbecher[e], Sándor Baran[d] and Zied Ben Bouallègue[e]
(e) ECMWF, Reading, United Kingdom
(d) University of Debrecen, Hungary

Introduction

- Ensemble development tends to use metrics for predictions of scalars, like the Continuous Ranked Probability Score, CRPS. Do we have a blind spot if we do not use measures that evaluate how well the relationships between different variables are predicted?
- Multivariate predictions can consist of a set of locations, different lead times, different variables.
- There are proper scores for multivariate predictions like the energy score and the logarithmic score.
- Here, recent work is summarised that focusses on a fair version of the logarithmic score and the extension of rank histograms to bivariate predictions.

The logarithmic score and ensemble size

Extend work of Siegert et al. (2019) to forecasts issued as **multivariate normal distributions**

The log score for multivariate normal distributions

Consider $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and a (regular) covariance matrix $\boldsymbol{\Sigma}$, representing an n-member forecast ensemble, and let $\boldsymbol{y}$ denote an observation.
The ensemble mean and the ensemble covariance matrix are

$$\boldsymbol{m} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \qquad \text{and} \qquad \mathbf{S} := \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^\top .$$

The scores for the distribution and the ensemble are

$$\mathrm{LogS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{y}) = \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\left(|\boldsymbol{\Sigma}|\right) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \quad \text{and}$$

$$\mathrm{LogS}(\boldsymbol{m}, \mathbf{S}; \boldsymbol{y}) = \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\left(|\mathbf{S}|\right) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{m})^\top \mathbf{S}^{-1}(\boldsymbol{y}-\boldsymbol{m}).$$

The fair logarithmic score for $\mathbb{R}^p$

$$\mathrm{LogS}_n^F(\boldsymbol{m}, \mathbf{S}; \boldsymbol{y}) = \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\left(|\mathbf{S}|\right) + \frac{n-p-2}{2(n-1)}(\boldsymbol{y}-\boldsymbol{m})^\top \mathbf{S}^{-1}(\boldsymbol{y}-\boldsymbol{m})$$
$$-\frac{1}{2}\left[\psi_p\left(\frac{n-1}{2}\right) - p\log\left(\frac{n-1}{2}\right) + \frac{p}{n}\right] \qquad \text{for } n > p+2.$$

The adjustments yield an estimate of the score of the distribution

$$\mathrm{E}\ \mathrm{LogS}_n^F(\boldsymbol{m}, \mathbf{S}; \boldsymbol{y}) = \mathrm{LogS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{y})$$

The derivation can be found in Leutbecher and Baran (2024). For scalars ($p=1$), LogS_n^F is identical to the result in Siegert et al. (2019).

2D rank histograms

Rank histograms are versatile tools that help assess the reliability of ensemble forecasts. While traditionally rank histograms are applied to univariate forecasts, they can also be used in a multivariate space. The proposed 2D ensemble rank histogram is a generalisation of the ensemble rank histogram to **bivariate ensemble forecasts**.

Methodology

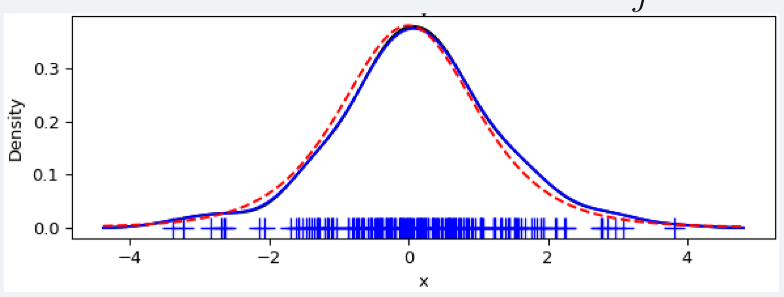
- The 2D vector composed of the ranks of the observation in the ensemble for the two components is used to determine the frequencies in the 2D rank histogram.
- In the univariate case, a flat rank histogram is interpreted as the ensemble being reliable (observations and ensemble members are statistically indistinguishable). For 2D rank histograms, the **ideal shape of a reliable ensemble is not known *a-priori***.
- Ensemble members can be used as pseudo-observations to build a reference 2D rank histogram for comparison. This reference is a representation of the **ensemble copula**.
- Let's recall Sklar's theorem (1959): a bivariate distribution F of the random variables v_1 and v_2 can be decomposed as

$$F(v_1, v_2) = \mathcal{C}\Big(F_{V_1}(v_1), F_{V_2}(v_2)\Big)$$

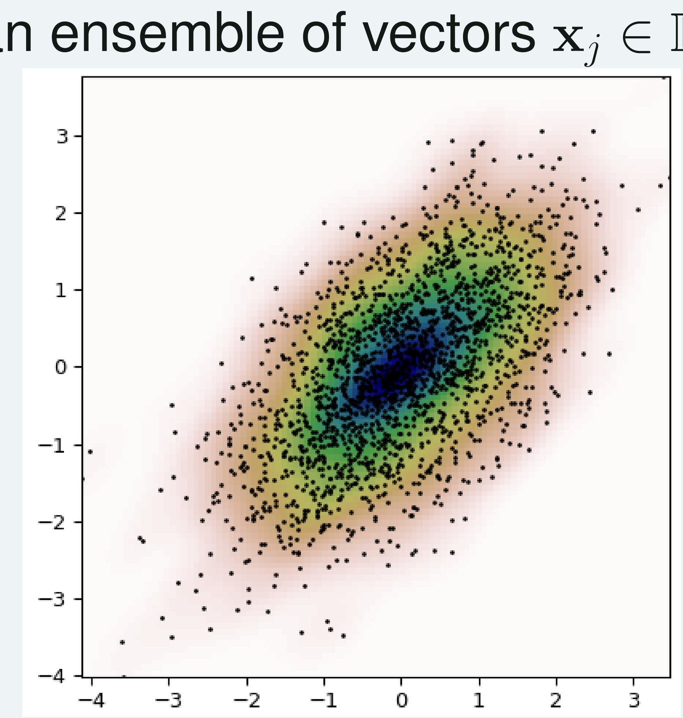
where F_{V_1} and F_{V_2} denote the univariate marginal distributions and $\mathcal{C}$ the **copula function** for the dependencies.

Univariate and multivariate ensemble verification

an ensemble of scalars $x_j \in \mathbb{R}$



an ensemble of vectors $\mathbf{x}_j \in \mathbb{R}^p$

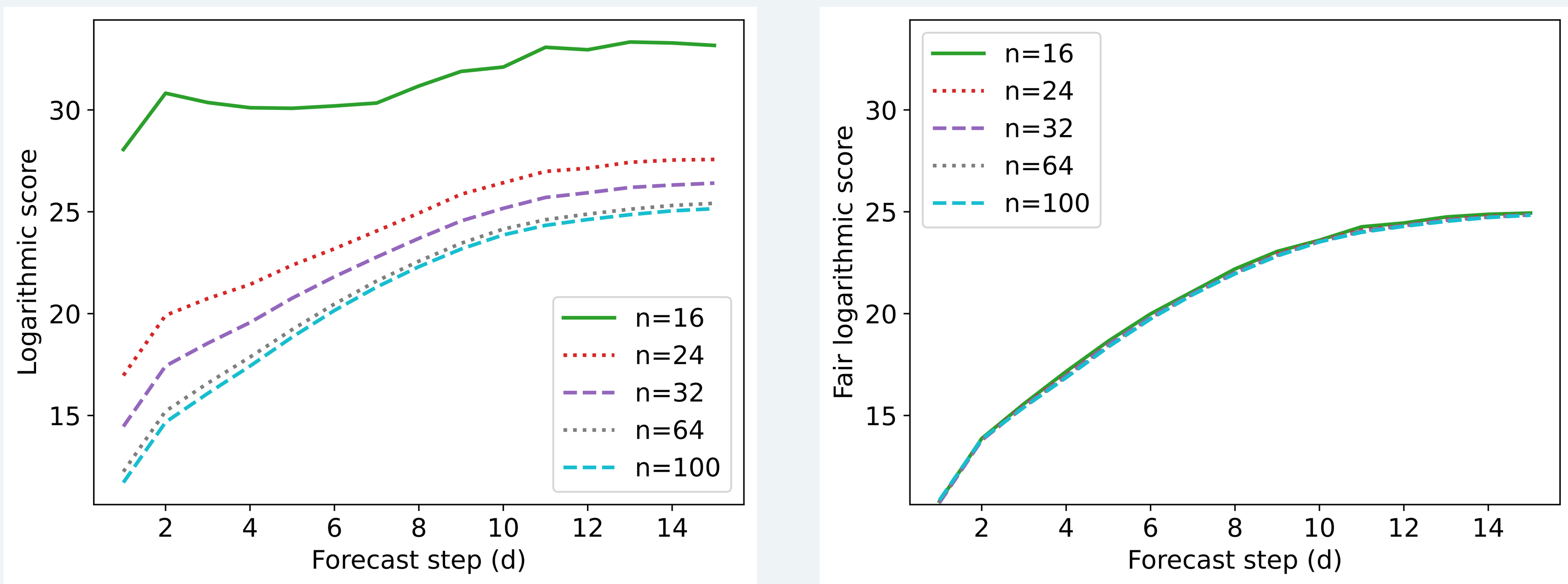


figures from scipy documentation, CC0 1.0

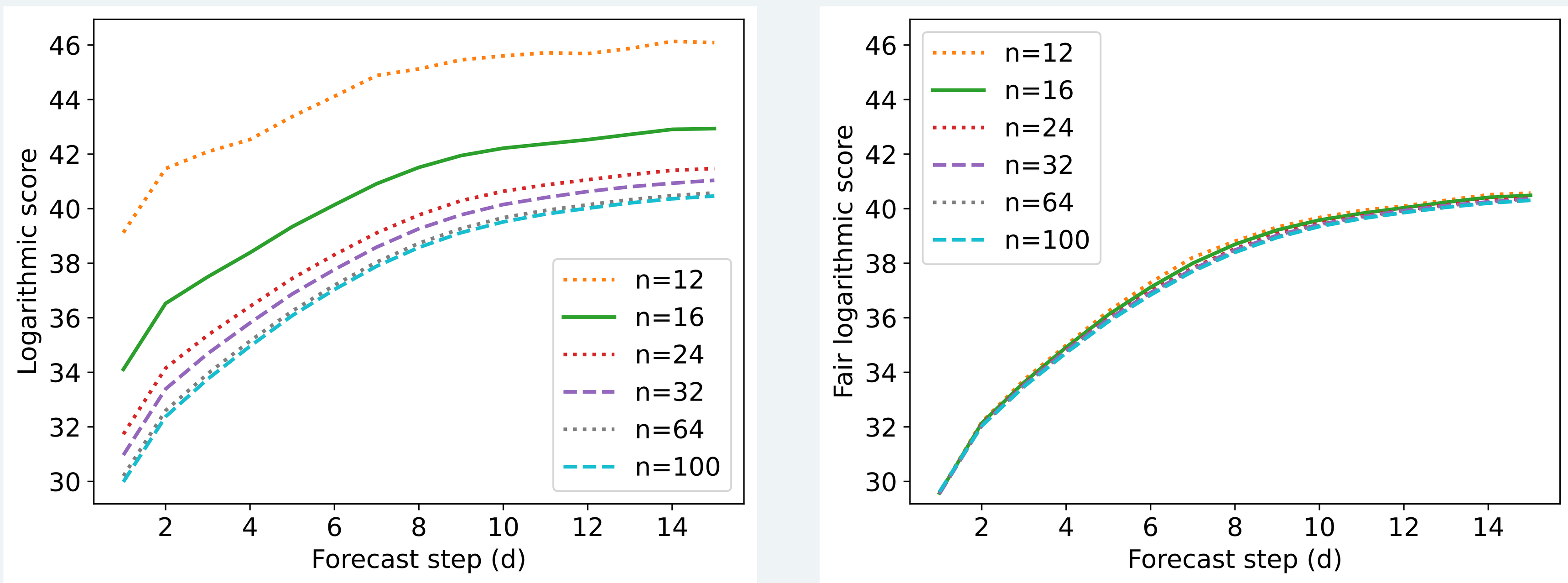
Results for 100-member subseasonal IFS ensemble

- Sep-Nov 2023, daily, 00 UTC, northern midlatitudes 35N–65N
- Scores for ensemble sizes $n = 12, 16, 24, \ldots, 100$

850 hPa temperature on 9-point stencil, $p = 9$



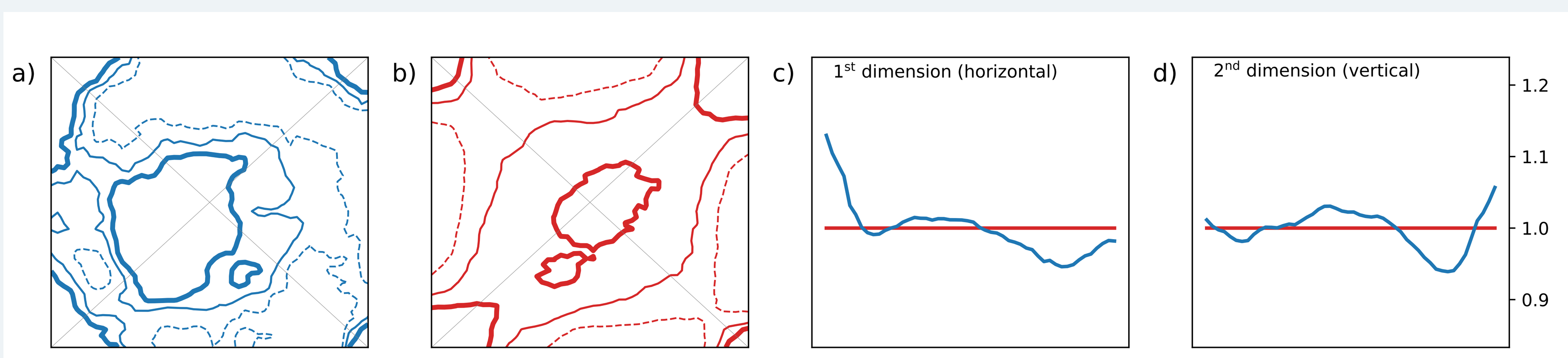
Geopotential on 6 pressure levels, $p = 6$



Results for the 50-member medium-range IFS ensemble

200 hPa horizontal wind components

Northern Hemisphere at Day 6



a) 2D rank histogram, b) ensemble copula, c) 1D rank histogram of the u-component, d) 1D rank histogram of the v-component. In panels (a) and (b), thin lines show the mean frequency, thick and dashed lines show deviations to the mean of $\pm$ half a standard deviation, respectively.

Whether the **dependencies between variables** in the ensemble reflect the dependencies in the observations can be assessed by comparing the 2D rank histogram with the ensemble copula: for a reliable ensemble, **(a) should look like (b)** within sampling uncertainty.

References

Leutbecher M, Baran S. 2024. Ensemble size dependence of the logarithmic score for forecasts issued as multivariate normal distributions. doi:10.48550/arXiv.2405.13400.

Siegert S, Ferro CAT, Stephenson DB, Leutbecher M. 2019. The ensemble-adjusted ignorance score for forecasts issued as normal distributions. *Quarterly Journal of the Royal Meteorological Society* **145**(S1): 129–139, doi:10.1002/qj.3447.